

INCENTIVES IN HIERARCHIES¹

Dilip Mookherjee²

August 26 2009

ABSTRACT

This chapter provides an overview of incentive-based theories of hierarchies. It starts by describing models focusing on supervisory functions of managers, addressing organizational scale diseconomies, and problems of collusion. Thereafter it examines costs and benefits of delegating decision-making authority, and related issues of monitoring and evaluation in a three layer hierarchy with a single owner and manager. Models of more complex hierarchies involving multiple layers, branches and different ways of grouping activities into departments are subsequently described. The chapter concludes with an assessment of the achievements and shortcomings of this literature, and possible directions for future research.

1. INTRODUCTION

The 20th century witnessed the emergence of large organizations in the private, public and non-profit sectors. Alfred Chandler's classic studies *Strategy and Structure* and *The Visible Hand* document the historical transformation of American industrial enterprises from small or medium sized owner-managed firms in the early 19th century into large business conglomerates controlled by a hierarchy of professional managers by the middle of the 20th century. Likewise government agencies managed by professional bureaucracies have grown in size and scope throughout the course of the 20th century, accompanying the growth of the role of governments in taxation, regulation, and delivery of public services.

The internal organization of private firms and government agencies is therefore a topic of considerable interest for both academics and policy practitioners interested in issues such as determinants of productivity, income distribution, human resource development, and efficiency of government. It is now commonly accepted that the size and 'internal structure' of firms is an important determinant of their productivity and profitability. As various management 'best-sellers' indicate, firms seem to vary considerably in how 'well-managed' they are. Organizations that are too large are widely believed to be overly bureaucratic, unresponsive and top management 'out of touch' with ground reality — and that this is the key problem with socialist economies with large state-owned firms or large government bureaucracies, compared with more decentralized market economies. Yet conventional economic theory has difficulty explaining why this is so, or why some

¹Preliminary version, prepared for *Handbook of Organizational Economics*, edited by Robert Gibbons and John Roberts, Princeton University Press, forthcoming. I thank Alberto Motta, Masatoshi Tsumagari and the editors for useful comments.

²Department of Economics, Boston University; dilipm@bu.edu

organizations are not as well-managed as others, or the exact manner in which size, structure and operating rules of organizations determine their effectiveness and productivity. Relevant attributes of organizational structure include the shape (e.g., the number of vertical layers, spans of control), allocation of authority and responsibility across managers, grouping of functions and managers into different departments, processes by which information is distributed, processed and communicated across managers, the way that managers are evaluated and compensated, and how all this eventually affects the way decisions get made and coordinated across the organization.

The internal organization of large firms is also a topic of considerable interest for those interested in inequality and social mobility, given the large fraction of the workforce in modern economies employed as ‘white-collar’ occupations involving managerial or secretarial or supervisory jobs. Theories of income distribution need to grapple with the question of what roles managers play in organizations, how they are evaluated and compensated, how these vary across different layers of management, as well as between ‘white-collar’ and ‘blue-collar’ workers.

Managerial hierarchies pose a significant challenge for conventional economic theory, raising issues such as the separation of ownership from management, and the functional role of managers in productive enterprises. What prevents owners from managing the firms they own? Alfred Marshall listed ‘management’ as a fourth critical factor of production, besides land, labor and capital. Yet more than a century later, models of management are still not part of the central corpus of the neoclassical theory of the firm. In the context of industrial or service enterprises that account for the bulk of all economic activity in developed countries, fixed factors such as land do not play an important role, creating challenges for explaining limits to the size of firms. While many believe that time and attention of top managers is one of the most important fixed factors that account for limits to the size of firms, modeling this has posed a significant challenge for economic theory. Part of the reason is the difficulty of modeling cognitive capacities of managers to process information, make decisions, and get things ‘done’ — that belong to the realm of ‘bounded rationality’ that Herbert Simon and Oliver Williamson have emphasized are key to understanding organizations and the role that managers play in them.

Nevertheless, the 19th century view of a firm as an entity whose sole objective is to maximize profits, has now evolved to a ‘nexus of contracts’ among a large number of important stakeholders and functionaries (see, e.g., Milgrom and Roberts (1992)). The firm is viewed as an organization or collective of different agents with dispersed information, responsibilities and non-congruent interests. Accordingly, problems of asymmetric information and incentives play an important role in the modern theory of the firm. Yet contemporary mainstream theory of information economics as represented by the principal-agent paradigm is still struggling to explain why authority and responsibility is distributed across different managers in an organization, instead of being centralized. The Revelation Principle which plays a central role in standard principal-agent theory (e.g., Myerson (1982)), states that any allocation of authority in an organization can be replicated by a degenerate centralized organization in which all agents communicate their private information to the owner or a central headquarter office (‘Principal’), which

processes this information and makes all relevant decisions, subsequently sends instructions to all agents concerning what they should do. In this kind of organization the only role of agents is to communicate information and receive instructions; the ‘mechanism’ needs to be designed in order to encourage honest and obedient behavior from all of them. In this setting, there is no role for a managerial hierarchy: all authority is vested in the Principal.

To create a theory of managerial organizations, it is necessary to go back to the foundational question: ‘what do managers do?’. How or why do they do these tasks better than the principal can? To quote Chandler on this:

“Just what, then, are the functions of the executives responsible for the functions of the enterprise? They coordinate, appraise and plan. They may, at the same time, do the actual buying, selling, advertising, accounting, manufacturing, engineering or research, but in the modern enterprise the execution or carrying out of these functions is usually left to such employees as salesmen, buyers, production supervisors and foremen, technicians and designers. In many cases, the executive does not even personally supervise the working force but rather administers the duties of other executives. In planning and coordinating the work of subordinate managers or supervisors, he allocates tasks and makes available the necessary equipment, materials, and other physical resources necessary to carry out the various jobs. In appraising their activities, he must decide whether the employees or subordinate managers are handling their tasks satisfactorily. If not, he can take action by changing or bringing in new physical equipment and supplies, by transferring or shifting the personnel, or by expanding or cutting down available funds. Thus, the term, *administration*, as used here, includes executive action and orders as well as the decisions taken in coordinating, appraising, and planning the work of the enterprise and in allocating its resources.” (Chandler (1962, p.8))

From a conceptual standpoint, therefore, there are at least three important roles that managers play: they (i) *process information*, helpful in decisions made by the firm regarding production, marketing, technology, employment and suppliers; (ii) *supervise*, i.e., generate and supply information used to evaluate and compensate employees and suppliers; and (iii) *make decisions* concerning production, marketing and contracting with employees and suppliers on behalf of the owners. The first two functions pertain to their role as processors or generators of information useful in making decisions. Only the last function includes actual decision-making. It is more comprehensive and includes some of the tasks described in (i) and (ii), such as processing of information generated by subordinates, supervising them, making decisions concerning employment, procurement and production within their respective ‘divisions’ or domains of control. Many organizations are organized hierarchically, i.e., distribute organizational tasks recursively across managers at successive layers. The Principal must be limited in her capacity to carry out these tasks, necessitating delegation of a substantial part of these to managers. The theory thus needs to model activities such as information processing, communication and

decision-making, and explain how these cognitive functions are efficiently distributed across owners and managers, their activities coordinated, evaluated and compensated.

This chapter provides an introduction to the literature on models of hierarchical organizations or contracting nexuses characterized by incentive problems (i.e., asymmetric information, non-congruent objectives and strategic behavior). Since theories of costly information processing are still in a state of relative infancy, and yet to be combined with incentive problems, the theories we will describe will focus on the second and third roles of managers described above, i.e., supervision and contracting with subordinates. The chapter by Garicano and van Zandt in this volume provides an overview of models which focus on the role of managers as processors of information *per se*. It also partially overlaps with the chapter on authority by Bolton and Dewatripont. The models we shall focus on will abstract from explicit consideration of costly information processing, even though this may constitute the underlying rationale for the existence of a managerial hierarchy in the first place.

In particular, we shall focus only on models in which incentive considerations play an important role. Accordingly many interesting models of hierarchies based on team theory (Radner and Marschak (1972)) which abstracts from incentive considerations will not be included. Even within incentive-based models of hierarchies, the coverage and emphasis will be biased in favor of those that I am personally familiar with, and help articulate some key results and the arguments underlying these.

Section 2 will describe the literature on ‘supervision hierarchies’ following the work of Calvo and Wellisz (1978), in which managers are modeled as supervisors. The rationale for supervision is that it helps limit the asymmetry of information between employees and owners, thus limiting the scope for opportunistic behavior by the former. Owing to limitations of time or expertise, owners may seek to supplement their own supervisory activities with information supplied by third party supervisors. Yet these hired supervisors also need to be motivated to apply effort in supervision, and thus need to be supervised themselves. This creates the rationale for a hierarchy of supervisors. Much of this literature has sought to explore implications for limits to the size of firms, as firms with a larger scale of operations employ more workers and thus need a larger hierarchy of supervisors, causing ‘organizational diseconomies of scale’ owing to resultant ‘losses of control’ across successive vertical layers of managers. This intuitive idea, expressed originally by Williamson (1967), however, has turned to be quite difficult to model. We also describe some recent literature on supervision hierarchies following Tirole (1986) which focuses on problems of collusion between supervisors and supervisees, rather than the question of scale diseconomies.

Section 3 turns to models focusing on the third, more comprehensive, function of managers as those delegated control by the Principal over decisions concerning production and contracting with subordinates. The class of models described in this section essentially study the cost or ‘control loss’ incurred by the Principal in delegating authority to managers over key productive and contracting decisions with employees and suppliers, owing to incentive problems (i.e., non-congruence of goals and information between owners and managers). The benefits of delegation are not studied in this approach, as

these are presumed to arise from unmodeled costs of centralizing all information processing and decision-making responsibilities with the Principal. The advantage of this approach is that it utilizes tools of mechanism design theory which rely on the assumption of ‘complete contracts’, in an environment where the Revelation Principle applies: the principal is assumed to be able to commit to a comprehensive contract; there are no costs of communication or contractual complexity; and agents behave noncooperatively. The later part of this Section describes a growing literature on models which extend this approach to contexts with collusion among agents.

Section 4 subsequently addresses the question of the benefits of delegation in conjunction with its costs, which necessitates a framework in which the Revelation Principle no longer applies and contracts are ‘incomplete’. We describe models based respectively on limited commitment ability of the Principal, and on costs of communication or complexity.

The preceding sections deal with the simplest possible model of a hierarchy involving three vertical layers, consisting of a single Principal at the top, a single manager in the middle, and one or two productive agents at the bottom. Models of larger and more complex hierarchies are considered in Section 5, which build on the simpler models in earlier sections. These focus on problems of horizontal coordination across departments, scale diseconomies and grouping of activities into departments (e.g., U-form versus M-form hierarchies).

Finally, Section 6 concludes with an assessment of the achievements and shortcomings of the literature, and possible research directions for the future.

2. ORGANIZATIONAL SCALE DISECONOMIES AND SUPERVISION HIERARCHIES

Williamson (1967) initiated the modern literature on limits to size of hierarchical firms arising from the ‘control losses’ associated with internal management problems. His theory is based on a ‘behavioral’ theory of how losses of information and control arise across vertical layers of management. The other ingredient in the theory are limits to the ‘span of control’ of any given manager: i.e., the number of subordinates that any manager can oversee owing to problems of limited time and attention. Given the limits of span of control, it is not possible for the firm to grow without adding vertical layers of management. But then the vertical control losses cascade across the various layers, implying rising inefficiency as the firm grows.

The source of these control losses are not explained by Williamson’s model. He provides as motivation ‘errors in serial reproduction’ demonstrated in social psychology experiments, wherein losses in information can occur when it passes through a succession of individuals, despite each individual in the chain being satisfied they have passed on all important features of their own information with little or no change. Hence top managers in organizations tend to be progressively out of touch with events at the ground level, as the number of vertical layers increase. Other sources of control loss are asymmetric information and opportunistic behavior by subordinates, which were not however part of the formal model.

The simplest version of Williamson’s model has an exogenously given span of control s at each layer. At the top level ($i = 1$) of the firm there is one owner or principal, supervising

s managers at level $i = 2$, each of whom are supervising s level-3 managers in turn, and so on. At the very bottom layer n , there are thus s^{n-1} employees, the production workers of the firm. All layers $i = 1, \dots, n - 1$ consist of managers or supervisors, layer i consisting of s^{i-1} managers. The production function yields the firm's output as a Cobb-Douglas function of the number of production workers s^{n-1} , and a level of total factor productivity that reflects a proportional loss of output owing to the control loss at each management layer: $y = (\bar{a}s)^{n-1}$, where $\bar{a} \in (0, 1)$ is the vertical control loss parameter reflecting communication and incentive problems.

Williamson also assumes an exogenous rate of wage progression across layers of the hierarchy. $\beta > 1$ is the ratio of wages at one level to the level below. Hence if w_0 denotes the wage of production workers, level $n - 1$ managers earn $\beta.w_0$, and level i managers earn $w_0.\beta^{n-i}$.

If the firm faces product price P on the output market (net of raw material costs per unit of output), and wage rate w_0 for production workers on the labor market, it selects a size (or number of layers n) to maximize $P(\bar{a}s)^{n-1} - w_0 \sum_{i=1}^n s^{i-1} \beta^{n-i}$. This yields an expression for the number of layers as a function of span of control, the vertical control loss parameter \bar{a} the wage rate for production workers, and the rate of wage growth across layers:

$$(1) \quad n^* = 1 + \frac{1}{\log \bar{a}} \left[\log \frac{w_0}{P} + \log \frac{s}{s - \beta} + \log \left(\frac{\log s}{\log \bar{a}s} \right) \right]$$

Firm size is finite if $\bar{a} < 1$, and tends to ∞ as control losses disappear $\bar{a} \rightarrow 1$. It is also increasing in s , the span of control.

An extended version of this model makes the span of control s endogenous by postulating that the vertical control loss at any layer is increasing in s . This captures the intuitive idea that limits on time and attention of any supervisor cause loss of control per subordinate, when the supervisor has to supervise more people. Hence \bar{a} is decreasing in s . With the specific functional form $\bar{a}(s) = \exp[-ks^2]$, Williamson shows that larger firms have higher n and lower s , but this result seems to depend on the specific functional form chosen.

Williamson's model has the virtue of delivering predictions concerning size and structure of firms based on technology, product and labor market characteristics that can be empirically tested. However, the key ingredients of the theory: vertical control loss (\bar{a}) and the rate of wage progression β in the hierarchy are taken as given. Much of the literature described in this chapter can be viewed as an attempt to endogenize these on the basis of incentive considerations, and thereby provide a micro-foundation for the Williamson model. As we shall see, this goal is yet to be fully achieved.

Calvo and Wellisz (1978) were among the first to do so, using a model of supervision and wage incentives. In this model, all workers and supervisors are identical: they have a utility function $u(c) - v(a)$ where u is a smooth, strictly increasing, concave function of consumption or earnings c , and v is a smooth, strictly increasing, convex function representing disutility of effort $a \in [0, 1]$ which is the fraction of the week worked, with $u(0) = v(0) = 0$. A production worker applying effort a generates revenue $P.a$ for the

employer, where P is a given parameter depending on output price, raw material costs and technology.

The supervision process is as follows. Supervisors allocate their time equally between monitoring all their subordinates. If the number of supervisors at level i is denoted by M_i , and if they work a_i fraction of the week, the amount of time devoted to supervision per level- $(i + 1)$ -employee is $\frac{a_i \cdot M_i}{M_{i+1}}$. With probability $p_{i+1} \equiv g(\frac{a_i \cdot M_i}{M_{i+1}})$ the true value of a_{i+1} , the time worked by any level- $(i + 1)$ employee will become known, where g is a strictly increasing function satisfying $g(0) = 0$ and $g(\infty) \in (0, 1]$. Here it is assumed that supervisors report honestly their findings to the employer.

Let w_{i+1} denote the wage paid for full-time work to a level- $(i + 1)$ employee. This will be determined endogenously, as explained below. If the employer learns from supervisor reports that an employee at level $i + 1$ worked a_{i+1} , this employee is paid $a_{i+1} \cdot w_{i+1}$. If no information is available, the employee is presumed to have worked fulltime, and paid a wage of w_{i+1} . Hence workers detected ‘shirking’ are punished by having wages withheld for the fraction of time they are reliably known to have ‘shirked’.

Consider a worker who is monitored with probability p and paid a fulltime wage of w . This worker will choose to work a^* fraction of the time, which maximizes $pu(w \cdot a) + (1 - p)u(w) - v(a)$. Clearly this is increasing in p and also a function of w : let this be denoted by $a^*(p; w)$. The worker will agree to work for the firm if $V(p, w) \equiv pu(wa^*(p, w)) + (1 - p)u(w) - v(a^*(p, w)) \geq \underline{u}$, a given positive level of utility that forms the worker’s outside option. Clearly $V(p, w)$ is increasing in w and decreasing in p .

Consider first a firm where the owner herself supervises all production workers, in what we may call a 1-layer hierarchy. Assume that the owner has a fixed amount of time, normalized to unity, available to supervise the workers. The optimal size of such a firm is defined by the solution to the following problem: choose n , the number of production workers, and a wage w , to maximize $nPa^*(p, w) - [pwa^*(p, w) + (1 - p)w]n$, subject to $p = g(\frac{1}{n})$ and the participation constraint for workers. It is easy to verify there is an interior optimum for both n and w for this problem. In particular, as n rises, the monitoring probability p falls, inducing workers to shirk. As n tends to ∞ , p tends to 0 and workers then tend to apply zero effort, in which case firm profits tend to $-\infty$ (since the firm needs to pay a positive wage to attract workers). Hence firm size is bounded if it owner-managed, i.e., a 1-layer hierarchy. Denote by π_1^* the maximized level of profit achieved by the owner in this case.

One way for the firm to expand is to hire third party supervisors. In a 2-layer hierarchy, the owner employs a number of supervisors and asks them to monitor production workers. The owner herself monitors the supervisors, which ensures that the supervisors exert effort in monitoring production workers, which in turns motivates the latter to apply effort in production. The incentive problem for choice of effort by supervisors as well as the participation constraint is exactly analogous to that of production workers. In a 2-layer hierarchy the owner selects the number of supervisors and the number of production workers supervised by each supervisor, as well as the wages at each layer. Hiring level-1 supervisors relieves the owner with regard to supervisory responsibility, thus enabling the firm to employ more production workers and ensure that their efforts do not get

diluted owing to the shrinking probability of being monitored when the owner alone has to carry out all the supervision. This is a formalization of the vertical ‘control loss’ which is increasing in the ‘span of control’. The Calvo-Wellisz model thus allows a simple and elegant model of these critical variables that were assumed exogenous in the Williamson model. Moreover, the wage structure in the hierarchy is also endogenously determined. The interesting question is whether these control losses arising for incentive reasons limit the size of the firm. Let π_n^* denote the level of maximum profit earned by the owner in a n -layer hierarchy.

The interesting result in the Calvo-Wellisz paper is that this model does **not** provide a limit to the size of the firm. The key Proposition is the following: *suppose $\pi_1^* > P$, i.e., it is more profitable for the owner to hire production workers and form a 1-layer hierarchy rather than do all the production work herself. Then $\pi_n^* \rightarrow \infty$ as $n \rightarrow \infty$.*

Here is an outline of the argument. It suffices to show there is a way of expanding the number of production workers indefinitely by raising n , which leads to unbounded profits. This involves replicating the optimal incentive arrangement for production workers for every layer of supervisor in a n -layer hierarchy. Specifically, letting a^*, w^*, n^* denote the optimal values of a, n, w in a 1-layer hierarchy, select a constant span of control $s^* = n^*$ at each level of the hierarchy, and pay the same wages at all levels. Hence the number of supervisors at pair of successive levels $i - 1$ and i are selected so that $\frac{a^* M_{i-1}}{M_i} = \frac{1}{n^*}$. Provided supervisors at level $i - 1$ are working a^* each, each level- i supervisor is monitored with the same probability $p^* \equiv \frac{1}{n^*}$ as production workers in a 1-layer hierarchy. Hence it is optimal for each such supervisor to agree to work for this firm, and work a^* fraction of the time. Letting W^* denote the expected wage cost per employee $p^* w^* a^* + (1 - p^*) w^*$, expanding the firm from $n - 1$ to n layers results in an increase in profit by $Pa^*(M_n - M_{n-1}) - M_n W^* = M_n [Pa^*(1 - \frac{M_{n-1}}{M_n}) - W^*] = M_n [Pa^* - P \frac{1}{n^*} - W^*] = \frac{M_n}{n^*} [(Pa^* - W^*)n^* - P] = \frac{M_n}{n^*} [\pi_1^* - P]$ which grows without limit as $M_n \rightarrow \infty$. Essentially, the mechanism in place in a 1-layer hierarchy can be replicated independently for each supervisor at each layer. Hence the model shows that incentive problems do not necessarily constitute a source of control loss that limits the size of firms. Formalizing the simple Williamsonian intuition will require some additional ingredients.

Of course the preceding model is somewhat special in a number of respects, such as the symmetry between workers and supervisors and their respective incentive problems. It turns out, however, that the basic result is robust to many variations that make the model more ‘realistic’ (Datta (1996), Tsumagari (1999)). Both these papers consider contexts where supervisory effort cannot be directly monitored. Datta studies a context where supervisors are evaluated indirectly by monitoring the output of production workers in their supervision. Here the Calvo-Wellisz replication arguments continue to apply under plausible assumptions. Tsumagari considers a context in which supervisors have overlapping jurisdictions, with cross-checking of reports of different supervisors. In this context the firm can expand horizontally rather than vertically, with each supervisor being cross-checked against reports of adjacent supervisors, and again firm size is unbounded under reasonable conditions.

Another variation was suggested by Calvo and Wellisz themselves, in which supervisees know in advance exactly when they are being monitored, so the fraction of time they work equals the fraction of time they are monitored. This implies that to get n production workers to work any fixed positive a fraction of the week, it is necessary to have them supervised by the same number of supervisors working a fraction of the week, and the same is true in turn for the supervisors themselves, so ultimately the owner will have to personally spend at least $a.n$ fraction of the week. Given the limit on the total time available to the owner, it is impossible to let n grow indefinitely while a is bounded away from zero. In that case firm size is limited. However, this is no longer so once there is some scope for random audits wherein workers do not know in advance when or whether they will be monitored. In that case one supervisor working a fraction of the week can supervise and ensure that $n > 1$ subordinates will also work a fraction of the week. The logic of the model described above then kicks in to ensure that the firm can grow indefinitely.

Qian (1994) returns to this question with a formulation that combines aspects of the Calvo-Wellisz model with Williamson's 1967 model. The critical departure from the Calvo-Wellisz model is the production function. Since the number of production workers in the n -layer Calvo-Wellisz hierarchy equals M_n , the gross revenue of the firm in the Calvo-Wellisz model equals $P.M_n.a_n$ if the layer- i employees apply effort a_i . Qian assumes instead a gross revenue which equals $P.M_n.a_n.a_{n-1} \dots a_1$. In other words, there is an additional source of revenue loss $a_{n-1} \dots a_1$ arising from existence of supervisors at intermediate levels of the hierarchy. Qian uses a variant of the incentive model of Calvo and Wellisz, whereby any shirking of the employee below the mandated effort level a^* is punished with a zero wage. Consequently, an employee monitored with probability p needs to be paid an efficiency wage of $\underline{w} \equiv \frac{v(a^*)}{p} = sv(a^*)$ in order to induce an effort of a^* , where $s \equiv \frac{1}{p}$ denotes the span of control of the employees' supervisor. The costs of increasing the span of control are thus represented by corresponding linear increase in the efficiency wage, for a given level of effort.

Using the formulation, Qian studies the problem of determining the 'optimal' hierarchy, i.e., number of layers, span of control at every layers, efforts and corresponding wages at each layer. If feasible effort a for any employee is zero or one, then clearly every employee must be induced to work ($a = 1$), and the formulation coincides with Calvo-Wellisz. In this case, there is no limit to firm size. The other case that Qian investigates is where a can be any number in the unit interval, and $v(a) \rightarrow \infty$ as $a \rightarrow 1$. In that case it is impossible to get anyone to work fulltime, and every agent must put in an interior $a \in (0, 1)$ in any optimal arrangement, with a bounded away from 1. Given the specification of the production function, however, this implies that the supplementary control losses $a_{n-1} \dots a_1$ arising due to the existence of intermediate supervisors must cascade, eventually implying that the firm's revenue converges to zero as the number of layers grows without bound. A limit to the size of the firm obtains. But it owes entirely to vertical control losses which remain unexplained. It is not clear why the effort of the supervisors have any effect over and above the way they influence the effort of those they supervise. Perhaps managers do other things apart from supervise workers, but these

are not included or explained. Hence the Calvo-Wellisz puzzle of limits to firm size does not really get resolved in this model.

2.1. Collusion and Supervision. Tirole (1986) pointed out an important qualitative difference between a firm where the owner personally supervises the worker, from one where she hires a third-party supervisor — the possibility that the supervisor and worker may collude. This is a factor ignored in the models of Calvo-Wellisz and Qian.

Tirole’s model has a Principal P and an agent A who produces revenue $\theta + a$ for P, where $\theta \in \{\theta_L, \theta_H\}$ is a productivity realization which is observed by A before choosing a level of effort a . Productivity can be either low (θ_L) or high (θ_H), where $\theta_H > \theta_L > 0$. P’s net payoff is the expected value of $\theta + a - w$, where w is the wage paid to A. A’s payoff is $u(w - v(a))$, where v is an increasing, convex function of $a \geq 0$, and u is a strictly increasing, concave function. A has a reservation utility normalized to 0.

P can hire a supervisor S, who may observe the realization of θ , though he cannot observe the agent’s effort a . The information that S can observe in the process of supervision is represented by a signal σ with three possible realizations: θ_L, θ_H and ϕ . If $\sigma = \theta_i, i = L, H$ the true state is exactly θ_i : in this case S learns the true state with certainty. However the signal may also equal ϕ in either state: in this case S does not learn the true state. Tirole assumes that the signal represents ‘hard’ information: S can submit definite evidence to P concerning the realization of the signal when it reveals the true state. S cannot fabricate evidence when there is none, i.e., cannot produce any evidence when $\sigma = \phi$. The only option for behaving strategically is to suppress evidence: S can claim $\sigma = \phi$ when he did receive evidence of the true state. Unlike the Calvo-Wellisz model, supervisory effort *per se* generates no disutility for S. The latter’s utility is $V(c)$ where c denotes the financial return of S and V is a strictly increasing, concave function. S also has a given outside option level of utility.

P lacks the capacity or time to supervise the agent herself. In this situation, she would design an incentive contract for A which, as per the standard model of incentives with adverse selection, motivate A to select the first-best effort a^* satisfying $v'(a^*) = 1$ when A observes $\theta = \theta_H$, and a lower level of effort \underline{a} when A observes $\theta = \theta_L$. In the former state A earns an incentive rent, i.e., obtains a wage high enough to end up with a payoff above his outside option. In the latter state A earns zero rent, i.e., a net payoff equal to the outside option. In this equilibrium, A is indifferent in the high productivity state between producing the revenue of $(\theta_H + a^*)$ and $(\theta_H + \underline{a})$. The rent earned by A in the high productivity state is the ‘price’ or ‘bonus’ paid by P to induce A to apply high effort in this state, rather than slacking off on effort (reducing a from a^* to $\underline{a} - (\theta_H - \theta_L)$) and pretending the state is θ_L instead of θ_H).

When S is hired and does not collude with A, he reports the realization of the signal truthfully to P. This reduces the informational asymmetry between P and A, enabling P to avoid paying the incentive rent to A when there is firm evidence that the true state is θ_H . This is the benefit P obtains from hiring S, which has to be traded off against the cost of hiring S. If S’s outside option is low enough, e.g., if it is zero, the cost of hiring S is low enough that it is profitable for P to hire S. Here S earns a fixed salary of c , just equal to the level necessary to induce him to accept the job.

But now A has an incentive to bribe S to suppress evidence of the high productivity state, in order to qualify for the incentive rent. S does not mind doing this, as the salary c is fixed, while A gains the incentive rent as a result of the suppression of evidence. Tirole assumes that there are no frictions in collusion — it can be modeled as an enforceable side-contract between S and A, which cannot be observed by P. Moreover, A knows exactly what evidence is available to S, so in the case where S receives evidence there is no asymmetric information between S and A. Effectively, then, S and A will then jointly behave as if they are a single player, earning a pecuniary equivalent payoff of $c + w - v(a)$ equal to the sum of their respective payoffs. The second-best payoff achieved by P is thus no longer achievable in the presence of collusion.

Tirole goes on to solve for the third-best contract, which maximizes P's expected payoff subject to both individual and joint collusive incentives of S and A. An argument analogous to the Revelation Principle shows that P can confine attention to contracts which are collusion-proof, i.e., in which S and A have no incentive to enter into a collusive agreement, and S reports his evidence truthfully to P. This is achieved partly by P offering S a selective 'reward' for disclosing evidence of high productivity, as well as reducing the 'power' of A's incentives and rents, both of which reduce the stakes for collusion. A has less to gain by bribing S to suppress evidence of high productivity, and S is less willing to suppress this evidence. The net consequence is that S will be forced to bear risk associated with the availability of evidence, for which P will have to pay S a risk premium if S is risk-averse. And A will be provided with less effort incentive. On both counts P's profitability will decline, compared with the second-best outcome. The extent to which this occurs depends on how risk averse S is. If S is risk-neutral, P does not lose at all: P can effectively 'sell' the firm to S and thereby overcome the problem of collusion entirely: there is no loss of either profitability or productivity. At the other extreme, if S is infinitely risk-averse no risk can be imposed on S at all, which implies no incentive rents can be paid to A; the firm ends up with low productivity and profitability. Nevertheless, even in this case, it pays P to hire S if the latter has a zero outside option, owing to the 'advocacy' role he plays on behalf of A (by providing evidence of state θ_L when it occurs, which is in the joint interest of S and A). In general, when S is somewhat but not infinitely risk-averse, the 3-layer hierarchy is more profitable than the 2-layer hierarchy in which S is not hired, though the agent may become less productive owing to the reduction in the 'power' of incentives.

Tirole's model provides an useful illustration of the problem of collusion, and its consequences. Worker incentives are lower in the three-layer hierarchy compared with the two-layer hierarchy. The 'costs' of collusion are the rents P has to pay the supervisor to deter collusion and the lower effort of the worker. The severity of these depend on the risk aversion of the supervisor. We shall see in subsequent sections that this result recurs in the literature on delegation. Kofman and Lawaree (1993) extend Tirole's model to include external and internal auditors, and explain the role of the former in cross-checking reports made by internal auditors who may be colluding with workers.

Does collusion serve as a source of control loss which limits the size of firms? Datta (1996) sketches an example where collusion limits firm size. Supervisors are evaluated by audits of production workers in their 'divisions'. Firms must grow by adding vertical layers

owing to endogenous limits on span of control, and collusion is deterred by inconsistency of audit reports at successive vertical layers. The larger the firm gets, the lower the likelihood of such inconsistencies becomes. On the other hand collusion does not limit the size of the firm in the Tsumagari (1999) model based on overlapping jurisdictions of different supervisors. In his model, the firm grows ‘horizontally’, whence there is no problem of cascading control losses.

In summary, models of supervision hierarchies have illustrated some of the difficulties in explaining limits to firm size. Collusion is probably important in understanding why larger organizations tend to be less productive and why middle managers earn large rents. But we still do not have tractable models of determinate firm size and structure based on collusion and supervision.

3. DELEGATION AND LOSS OF CONTROL WITH COMPLETE CONTRACTS

The previous section has focused on the supervisory role of managers, pertaining to generation of information concerning the activities of subordinates which helps evaluate and motivate the latter. Supervision *per se* does not include any responsibility for actual decisions. Conceptually there is little distinction in this role from those of external consultants that the principal may rely on for information relevant to decision-making. In the supervision models, the principal still contracts with all workers and suppliers personally, and makes all relevant production, technology, marketing and sourcing decisions.

In this and succeeding sections we focus on the question of *delegation* of these key decisions by the principal to managers. As the quote from Chandler in the Introduction indicates, the managers may play some role in the production process themselves, but at higher levels of the hierarchy they rarely do. Their functions include contracting and supervision of subordinates, allocating resources and coordinating activities of departments under their control, apart from long-term planning. The principal or owners rarely get personally involved in the internal administration of the enterprise.

From an analytical perspective, we therefore need to understand the implications of delegation of contracting with subordinates to managers, where the term ‘contracting’ includes production and sourcing decisions, coordination of production and allocation of resources, recruitment, supervision and evaluation of employees and suppliers. In this section we shall not pose the question why the principal cannot carry out these tasks herself, i.e., what the value of delegation of contracting is. Instead, we shall examine the incentive problem inherent in delegation, where the manager is self-interested and may use his delegated authority in his own interest rather than the principal’s. This is precisely the problem of ‘loss of control’. The question posed is: what are the determinants of this loss of control? What are strategies for ameliorating it? How should managers be evaluated and compensated?

In the first subsection below, we shall assume the conditions underlying the Revelation Principle are valid: there are no costs of communication, complexity or information processing; the principal can commit to a comprehensive contract; and agents do not collude. Under these circumstances the Revelation Principle (e.g., Myerson (1982)) tells us that centralized decision making — where the Principal contracts and communicated

with all agents personally, and makes all relevant decisions based on reports of private information of the agents — can always perform at least as well as any method of decentralizing these responsibilities. Hence delegation can never outperform centralization. The question addressed instead is: can delegation of key responsibilities to self-interested managers achieve the same level of profit for the Principal as any centralized mechanism? If so, delegation is costless, and is a way of implementing the optimal mechanism.

Subsequent to this, we shall explore the consequences of dropping different assumptions underlying the Revelation Principle. The latter part of this section will explore the effects of collusion among agents, preserving all the other assumptions. The following Section will then examine the consequences of inability of the Principal to commit to a mechanism, or of costs of communication and complexity.

The ‘standard’ model in this Section will be very simple, amounting to an adverse selection contracting problem. Analogous questions in a moral hazard setting have been studied by Baliga and Sjostrom (1998) and Macho-Stadler and Perez-Castrillo (1998). There is one Principal (P), one or two agents (A, or $A_i, i = 1, 2$), and a third-party manager (M). Extension of these models to incorporate more agents, managers or layers will be covered in the penultimate section. The agents carry out productive tasks which generate a revenue or gross benefit B for P. In the single agent case, A delivers a good or service $a \geq 0$ which generates a benefit $B(a)$. In the two agent case A_i delivers $a_i \geq 0$ and the benefit is $B(a_1, a_2)$. The two goods delivered could be perfect substitutes in which case B depends only on $a_1 + a_2$: in this case the mechanism design problem reduces to a procurement auction with variable quantities. Or the two goods could be perfect complements: $B = \min\{a_1, a_2\}$. More generally, a_1 and a_2 could be neither perfect substitutes nor perfect complements, as represented by a CES production function.

Agents are privately informed regarding their cost of production, before they need to commit to deciding whether or not to participate in the mechanisms concerned. In other words, contracting occurs at the *interim* stage, where each agent knows his own cost but not the cost realizations of others. Most authors assume that each agent’s unit cost does not vary with the level of production, and this unit cost is uncertain. Some authors assume these costs can take one of two possible values; others assume they are distributed on an interval of the real line, with a standard monotone hazard rate assumption on the distribution which ensures that global incentive constraints can be ignored. Specifically, if θ_i is the unit cost incurred by A_i , it is distributed according to a c.d.f. F_i and density f_i on an interval $[\underline{\theta}_i, \bar{\theta}_i]$ where $\theta_i + \frac{F_i(\theta_i)}{f_i(\theta_i)}$ is nondecreasing in θ_i , an assumption satisfied by the uniform, exponential and many other well-known distributions. Throughout it is assumed that θ_1, θ_2 are independently distributed. Moreover, most formulations assume risk-neutrality and zero outside option utilities for agents. Hence P’s payoff is $B(a_1, a_2) - t_1 - t_2 - t_M$ where t_1, t_2, t_M denotes transfers made by P to A_1, A_2, M respectively. The payoff of A_i is $T_i - \theta_i a_i$, and of M is T_M , where T_i, T_M denote net transfers received by A_i, M respectively. In the centralized mechanism without any collusion, $T_i = t_i, T_M = t_M$, but this is not the case in decentralized mechanisms or in the presence of collusion owing to side-transfers between the agents.

Many authors have studied what we shall call the *restricted* version of the standard problem, in which each agent's cost takes one of two possible values, and the inputs produced by the two agents are perfect complements. Other authors have studied the *continuous* version of this problem, in which costs are distributed according to a density on an interval on the real line, and no restriction is imposed on the benefit function.

Decision-making responsibility can be delegated either to one of the productive agents, or to the third-party manager M . The latter may also act as a supervisor of the productive agents, receiving a private signal of their cost realizations. Agents and the manager all act in self-interested fashion.

In this section we assume that it is possible for anyone with contracting authority to commit to a comprehensive contract, which includes details of communication protocols (i.e., the message sets, who communicates with whom, when etc.), production assignments and transfers. Owing to the Revelation Principle it will suffice to focus on revelation mechanisms in the centralized regime, where agents report their private information, and receive instructions concerning production assignments. The outputs they deliver a_i are assumed to be costlessly monitored; transfers can be conditioned on these, and/or reports communicated. In the decentralized regime, assumptions concerning sequencing, communication and observability will be made explicit in due course.

In the setting with one agent, the following mechanisms can be compared: (a) centralization, denoted $(P^* - A)$, where the $*$ depicts the decision-maker and a $-$ indicates a line of hierarchical control; (b) agent-based-decentralization, denoted $(P - A^*)$; and (c) manager-based-decentralization, denoted $(P - M^* - A)$. In $(P^* - A)$, P retains all decision rights. In $(P - A^*)$, P contracts with A but delegates production decisions to A . In $(P - M^* - A)$, P delegates contracting and production decisions to M , and contracts with M alone. M contracts in turn with A . M is evaluated and compensated by P based on the revenue generated. Other mechanisms are also possible, such as $(P^* - (M, A))$, where M 's role is purely supervisory or consultative, thus is on the same hierarchical level as A , depicted by the parentheses $()$ around the set of agents M, A at that level. Here P contracts with both M and A , and M 's responsibility is limited to submitting a report to P concerning the signal concerning A 's cost. P uses this information to evaluate and compensate A .

In the setting with two agents, the corresponding mechanisms are: (a) centralization $(P^* - (A_1, A_2))$, (b) agent-based-delegation $(P - A_1^* - A_2)$ where P delegates to A_1 the responsibility of communicating and contracting with A_2 , and (c) manager-based-delegation $(P - M^* - (A_1, A_2))$ where P delegates to M the responsibility of contracting with A_1, A_2 . Variants of these include $(P - (A_1^*, A_2^*))$, where P contracts with the two agents, but delegates to them the right to make their own production decisions. Another is $(P^* - (M, A_1, A_2))$ where P retains all decision rights and the sole responsibility of M is to communicate reports to P .

3.1. Single Agent; No Collusion. Here A produces a at cost $\theta.a$, P 's net payoff is $B(a) - t$ and A 's is $t - \theta.q$. It is immediate that *centralization* $(P^* - A)$ and *agent-based decentralization* $(P - A^*)$ achieve equivalent outcomes, in what is often referred to as the *Taxation Principle*. Under the former, P designs an incentive compatible revelation

mechanism $a(\theta), t(\theta)$, in which A is always motivated to report truthfully. Equivalent outcomes are achieved in agent-based decentralization by P offering A instead a nonlinear incentive mechanism $t(a)$ equal to $t(\theta)$ if $a = a(\theta)$ for some θ , and 0 otherwise. The incentive compatibility of the original revelation mechanism implies that this nonlinear function is well-defined, and motivates A to select the same production in every state as P did in the centralized regime. The converse is equally straightforward, amounting to the proof of the Revelation Principle. Generalizations of this result to contexts with production or monitoring uncertainty (i.e., where P's revenue or measured performance is not a but some function $R(a, \epsilon)$ where ϵ is a random variable unobserved by A at the time of selecting a) have been provided by Melumad and Reichelstein (1987).

When P has the opportunity of using the services of M to supervise the agent, and M and A do not collude, P can obtain M's information costlessly in the centralized mechanism $P^* - (M, A)$, since M does not have any incentive to hide his information. Here the outcome is the same as if P contracted with A with the same information as M. This is no longer the case in the decentralized variant $(P - M^* - A)$, since M can act strategically with respect to his information. This is easy to see if M is perfectly informed about the cost realization θ : in that case M can procure any quantity a from A at cost $\theta.a$, and the problem is equivalent to P contracting with A in the absence of M. The same result is also evident when M's information about θ is identical to P's. Hence delegation to M is typically costly in this setting.

3.2. Two Agents, No Collusion. Consider the comparison between the centralized $(P^* - (A_1, A_2))$ and decentralized $(P - A_1^* - A_2)$ variants, a question which has been studied by a number of different authors (Baron and Besanko (1992), Gilbert and Riordan (1995), Melumad, Mookherjee and Reichelstein (1992, 1995), Laffont and Martimort (1998), Severinov (2008)). In the former, the Revelation Principle implies P can restrict attention to incentive compatible revelation mechanisms $\{a_i(\theta_1, \theta_2), t_i(\theta_1, \theta_2)\}_{i=1,2}$, where the two agents submit cost reports simultaneously and are motivated to report truthfully. The optimal centralized mechanism can be solved using standard techniques of auction theory, for the continuous version of the standard model. Let $h_i(\theta_i) \equiv \theta_i + \frac{F_i(\theta_i)}{f_i(\theta_i)}$ denote A_i 's 'virtual' cost. Then the optimal production assignments $\{a_i^*(\theta_1, \theta_2)\}_{i=1,2}$ must be chosen to maximize

$$(2) \quad B(a_1, a_2) - \sum_{i=1}^2 h_i(\theta_i) a_i.$$

The optimal transfers are given by

$$(3) \quad t_i^*(\theta_1, \theta_2) = \theta_i a_i^*(\theta_1, \theta_2) + \int_{\theta_i}^{\bar{\theta}_i} a_i^*(y, \theta_j) dy$$

where the second term on the right-hand-side represents the incentive rent paid to A_i , which is decreasing in θ_i and is zero for the highest cost type $\bar{\theta}_i$.

There are many variants of the decentralized version, differing with respect to the precise sequencing of contracting, and of the information available to P concerning the 'performance' of the manager-cum-worker A_1 .

A key result here is that *delegation to A_1 is optimal, i.e., can achieve the same expected profit as centralized contracting, if (a) the sequence of contracting is ‘top-down’ in the sense that P contracts with A_1 ³, before A_1 communicates or contracts with A_2 , (b) P monitors both the gross revenue B and **either** the payment t_2 made by A_1 to A_2 **or** the input a_1 delivered by A_1 alone; and (c) A_1 is risk-neutral and not subject to any limited liability constraint.*⁴

Here is an outline of the argument.⁵ P can offer the following contract of the following form to A_1 : pay

$$(4) \quad t_1(B, a_1; \theta_1) = B - [h_1(\theta_1) - \theta_1]a_1 + \beta(\theta_1)$$

which is interpreted as follows. The manager is paid a salary β which depends on a report of his own cost, plus a performance bonus equal to the revenue delivered to P , less a ‘tax’ based on A_1 ’s own contribution a_1 at the rate of the informational rent $[h_1(\theta_1) - \theta_1]$ corresponding to the cost report θ_1 . If A_1 has decided to participate in this contract and reported his own cost truthfully, he will subsequently choose a contract $t_2(\theta_2|\theta_1), a_2(\theta_2|\theta_1)$ for A_2 and make his own production decision $a_1(\theta_2|\theta_1)$ in order to maximize the expected value of $B(a_1, a_2) - [h_1(\theta_1) - \theta_1]a_1 - t_2 - \theta_1 a_1$, subject to incentive and participation constraints for A_2 . Again, this is a standard contracting problem with a single agent, and is solved as follows: the production assignments are selected to maximize (pointwise at each state (θ_1, θ_2)): the value of $B(a_1, a_2) - [h_1(\theta_1) - \theta_1]a_1 - h_2(\theta_2)a_2 - \theta_1 a_1$, which reduces exactly to expression (2) for the objective function of P in the optimal centralized contracting problem. Hence A_1 will select the same production assignments that P would select under centralization. It is also evident that he will also choose the same transfers for A_2 .

The rest of the argument involves showing that the $\beta(\theta_1)$ function can be chosen so that A_1 also ends up with the same transfers, and has an incentive to report his own cost truthfully to P at the first stage. Here the sequencing assumption (a) and the absence of limited liability assumption (c) play an important role, as we explain below.

The key problem with delegation is the potential for *double marginalization of rents (DMR)*, well-known in the IO literature as the main argument in favor of vertical integration (see, e.g., Tirole (1988)). Contracting with adverse selection gives rise to ‘monopsony’ distortions: owing to the trade-off between productive efficiency and incentive rents, too little tends to be procured from a first-best standpoint. Centralization involves only one layer of contracting and accordingly one layer of such ‘under-sourcing’. Delegation involves two successive layers of contracting, with two successive rounds of ‘under-sourcing’. The extra distortion arises from the monopsony power awarded to A_1 in contracting with A_2 , which induces A_1 to ‘out-source’ too little to A_2 .

³This includes both a commitment from A_1 to participate in the mechanism, as well as submission of a report concerning his own state.

⁴Severinov (2008) shows that additional conditions on the degree of complementarity between the inputs supplied by the two agents are needed if cost types have a discrete rather than a continuous distribution.

⁵See Melumad, Mookherjee and Reichelstein (1995, Theorem 1) for further details.

This problem is avoided if P can disaggregate the allocation of aggregate revenue between A_1 's contribution and A_2 's contribution. In the delegation mechanism discussed above, P imposes a 'tax' $[h_1(\theta_1) - \theta_1]a_1$ on self-production by A_1 , which is equivalent to a subsidy on 'out-sourcing' to A_2 . This corrects for the tendency for delegation to involve an additional layer of monopsony distortion.

A similar outcome can be ensured if P can monitor the payment t_2 made by A_1 to A_2 . In that case P can compensate A_1 as follows:

$$(5) \quad t_1 = \gamma_1(\theta_1)[B - t_2] + \beta(\theta_1)$$

where $\gamma_1(\theta_1) \equiv [1 + \frac{F_1(\theta_1)}{\theta_1 f_1(\theta_1)}]^{-1}$ is a 'bonus coefficient' applied to a measure of 'profit' (equal to revenue B minus 'cost' t_2) of the 'division' or 'profit center' managed by A_1 . Melumad, Mookherjee and Reichelstein (1992) use this as a model for 'responsibility centers' within firms. If there are more agents A_3, A_4, \dots at the third layer, the same result goes through with a divisional profit measure $B - t_2 - t_3 - t_4 \dots$ where cost incurred with respect to different employees and suppliers are linearly aggregated to form a divisional cost which is subtracted from revenue. The use of a bonus coefficient γ_1 which lies between 0 and 1 effectively subsidizes 'out-sourcing' to subordinates. Observability of aggregate cost of the division suffices to overcome the DMR problem. It is not necessary for P to observe further details of the side-contracts designed by A_1 , nor the communication of the latter with his subordinates.

Nevertheless, *some* monitoring of side-contracts (or A_1 's contribution a_1) is essential, as long as there is some substitutability between a_1 and a_2 . Only in the case of perfect complementarity (a Leontief production function) can the DMR problem be avoided if P does not monitor any aspect of the transactions between A_1 and his subordinates. And even in that case, there would be a tendency for A_1 to generate 'too much' aggregate revenue. This provides a possible explanation for the 'empire-building' tendencies of managers, which shareholders seek to limit in various ways (such as using a compensation formula such as (5) where the fractional bonus coefficient reduces the manager's inclination to over-expand operations). The production level desired by P should maximize (2), which reduces to $B(q) - [h_1(\theta_1) + h_2(\theta_2)]q$ in the case of perfect complementarity ($a_1 = a_2 = q$). Hence the 'desired' measure of unit cost of producing q is $h_1(\theta_1) + h_2(\theta_2)$, whereas A_1 would personally incur a unit cost of $\theta_1 + h_2(\theta_2)$ when delegated responsibility for making production decisions. The conflict of interest between owner and manager arises owing to the informational rents earned by the manager, which the manager sees as a benefit but the shareholders see as a cost. Baron and Besanko (1992) who study the case of perfect complementarity avoid this problem by centralizing the production decision while delegating A_1 only the responsibility for contracting and communicating with A_2 . The DMR problem is then avoided by reducing the extent of responsibility delegated.

The other assumptions (a) and (c) also play an essential role in ensuring the optimality of delegation. Delegation enlarges the extent of asymmetric information between P and A_1 , as the latter is now privately informed both about his own cost θ_1 as well as that of his subordinates θ_2 . This would expand the informational rents that A_1 would earn *vis-a-vis* P, on top of the rents that A_2 earns. This is another version of the DMR problem. It is avoided with assumption (a) of top-down contracting, since A_1 contracts with P *before*

he learns the realization of θ_2 . These additional rents can then be ‘taxed’ away at the time of contracting. For this it is necessary that A_1 end up with a negative payoff in some states of the world (e.g., when both θ_1 and θ_2 end up too high). This would not be possible if there were a limited liability constraint requiring A_1 ’s payoff be nonnegative in all states of the world. This is the central point made by McAfee and McMillan (1995), which underlies their model of organizational diseconomies of scale. Such a context is similar to one where A_1 communicates with A_2 before contracting with P, so the former knows the state of the world before committing to participate in the contract offered by P. In that world, A_1 will earn additional informational rents owing to the privacy of his information concerning the realization of θ_2 . His perceived unit cost of production in the case of perfect complementarity will be $[\theta_1 + h_2(\theta_2)]$, and so will earn a rent associated with knowledge of the second component $h_2(\theta_2)$ as well as the first θ_1 . There is then a ‘cascading of information rents’ across successive vertical layers of the hierarchy.

Similarly, if A_1 were risk-averse, he would incur risk associated with realization of his subordinates’ costs in the delegation arrangement, for which P would have to compensate A_1 with an additional risk premium. This is the main idea underlying Faure-Grimaud and Martimort (2001) and Faure-Grimaud, Laffont and Martimort (2000)’s model of the cost of delegation. The point is qualitatively similar to that made by Tirole (1986) in a different setting, as well as McAfee and McMillan (1995). It is particularly relevant to the case of managers in firms who are likely to be risk-averse and subject to limited liability concerns. The optimality result seems applicable, if at all, to the case of supply chains in defense procurement or health-care contracting in which P is the government and the agents correspond to large corporations with deep pockets and ample access to capital markets.

Nevertheless it is important to note that the result concerning optimality of delegation should not be interpreted as solving the problem by ‘selling the firm’ to A_1 . This is evident, for instance, in the ‘profit center’ arrangement described by the incentive scheme (5), where the use of the fractional bonus coefficient γ_1 implies that A_1 is not a full residual claimant on the firm’s profit net of his own compensation. P shares in the firm’s profits with the manager. Such profit-sharing is essential to ensure that managers internalize the objectives of owners.

3.3. Delegation to Information Intermediaries. Consider next the costs of delegating management of the firm with two (or more) productive workers to a third-party supervisor or intermediary M who does not carry out any productive tasks. As Chandler noted, most ‘top-level’ executives in large organizations are not involved in any productive activity themselves. The value of hiring such professional managers lies presumably in their ability to supervise and coordinate the activities of productive agents. So suppose that M receives signals about the realization of costs of different agents. Presumably, M is better informed than P, otherwise it would not pay for P to hire the former.

An interesting question here are the costs and benefits of P managing the firm herself (the $(P^* - (A_1, A_2))$ organization) with delegating management to M (the $(P - M^* - (A_1, A_2))$

organization).⁶ The advantage of delegation is that it takes advantage of M 's expertise and supervision. The disadvantage is that M will behave strategically with respect to this information *vis-a-vis* P and extract corresponding informational rents.

It is simplest to consider the case where M is perfectly informed about the costs of the two agents. Then the delegated organization reduces to the problem of P contracting with a single consolidated agent who produces both inputs (a_1, a_2) at a cost of $\theta_1 a_1 + \theta_2 a_2$, since this will be the cost that M will incur in procuring these inputs from the agents. Hence the comparison is between P contracting with two separate agents each of whom produces one of the inputs, with a consolidated agent who produces both at the same aggregate cost.

This problem has been considered by Baron and Besanko (1992), Gilbert and Riordan (1995), Mookherjee and Tsumagari (2004) and Severinov (2008). The main result is the following. *If M is perfectly informed about agents' costs, delegation to M is better than P managing the firm herself without involving M , if the two inputs a_1, a_2 are complements in the sense that $\frac{\partial^2 B}{\partial a_1 \partial a_2} \geq 0$ and the costs have i.i.d. exponential distributions with a lower bound of 0. The converse is true if a_1, a_2 are perfect or near-perfect substitutes (i.e., infinite or near-infinite elasticity of substitution).*⁷

The intuitive explanation of this is as follows. When P contracts with separate suppliers, each supplier exerts an externality on the other supplier through their cost reports. This gets internalized when the suppliers are consolidated into a single entity. If the inputs are near-perfect substitutes, the two agents are competing suppliers, and a lower cost report by one agent tends to decrease the production target awarded to and hence the payoff earned by the other agent. Internalizing this externality implies that cost reports will tend to become higher — which lowers P 's profits. Consolidation here suppresses competition. On the other hand when the two inputs are complements, a lower cost report by one agent expands the production target and the payoff of the other agent. Internalization of this externality results in lower cost reports — which works to P 's advantage. Here consolidation fosters 'cooperation'.

There is another effect of consolidation: two one-dimensional adverse selection problems are replaced by a single multi-dimensional problem. The consolidated agent can consider 'coordinated deviations' of cost reports along the two dimensions that were not possible

⁶There is a third alternative ($P^* - (M, A_1, A_2)$) in which P retains control and treats M as a supervisor rather than a manager. In the absence of collusion, P can costlessly acquire M 's information and then use this to contract with the agents. However, for the reasons provided by Tirole (1986), this would be vulnerable to collusion between M and the agents. In the presence of collusion, which we consider in the next subsection, it will turn out that this alternative is closely related to the one we are considering now, where authority is delegated to M .

⁷Again, this is for the case of continuously distributed costs. The result as stated above is an implication of Proposition 5 in Mookherjee and Tsumagari (2004). Severinov (2008) considers the case of a discrete set of cost types, and provides detailed results concerning how the comparison depends on the degree of substitutability and asymmetry across agents. However, the results are broadly similar to the continuous distribution case.

when the agents were separate. This adds to the problems of control of M . These problems are not severe when the two agents are *ex ante* symmetric and have exponentially distributed costs (which contains uniformly distributed cost as a special case).⁸

3.4. Delegation and Collusion. How do the preceding results get modified in the presence of collusion among agents? One heuristic view is that centralized contracting is generally vulnerable to collusion (e.g., it undermines P's efforts to induce competition among agents when they supply substitutes), while delegation is not vulnerable as the latter already incorporates a form of side-contracting between agents. Hence collusion increases the relative value of delegation, and renders it optimal in a wider class of circumstances. Extending the models considered above allows us to appraise the extent to which this view is correct. In fact, once we allow for collusion among agents, the Revelation Principle no longer applies. Is it even possible that delegation can then be superior to centralization?

The first note of caution is that the results concerning the incentive-constrained optimality of delegation ($P - A_1^* - A_2$) that was described above, relied on particular assumptions regarding the nature of side-contracting which are inconsistent with collusive behavior. For instance, P must be able to observe side payments between the two agents, or their relative contribution to the firm's revenues. The 'profit center' arrangement implies A_1 's compensation increases by less than a dollar for every dollar's increment in payments to A_2 . This will tempt A_1 to "pad" costs or payments to A_2 , in exchange for 'under-the-table' kickbacks. The manager must be prevented from communicating with the subordinate before contracting with the owner. Preventing either of these will typically be difficult for P. In the absence of ability to monitor side-payments between the agents, we have already noted that the firm will be prone to the 'empire-building' tendency of the manager, and delegation is then no longer second-best.

However, neither is centralization invulnerable to collusion. For instance, consider the case of a procurement auction designed by P, between two competing and *ex ante* symmetric agents. The second-best mechanism is a second-price auction, which creates strong incentives for the suppliers to enter into a hidden agreement to raise their cost reports. Whether centralization or delegation is more vulnerable to collusion is therefore a nontrivial question.

The answer depends partly on the precise way that collusion is modeled. Most of the literature follows Tirole (1986) and ignores possible problems with the enforceability of side-contracts. The usual 'handwaving' justification offered for this assumption is that the agents are engaged in a long-term relationship with one another and can thus enter into self-enforcing agreements. Additional problems arise with the allocation of bargaining power within the coalition of colluding agents. Owing to asymmetric information among the agents regarding their respective costs, the Coase Theorem does not apply to the analysis of *ex ante* optimal side contracts: the allocation of bargaining power affects decisions they make concerning the cost reports they decide to send P in a coordinated fashion. This is one complicating feature of this class of models compared

⁸Da Rocha and de Frutos (1999) and Severinov (2008) provide examples where consolidation performs worse owing to asymmetries between the agents, even when they perform complementary tasks.

with the context studied by Tirole (where there was effectively complete information within the coalition of the agent and the supervisor, so that their joint behavior could be represented simply by that of a consolidated entity earning the sum of their respective payoffs). There is a nontrivial contracting-within-contracting aspect to the problem.

A related question concerns how the allocation of bargaining power between colluding agents varies between the centralized and decentralized contracting regimes. Laffont and Martimort (1998) postulate that it is symmetric in the centralized regime, and asymmetric in the decentralized regime. This seems intuitive, insofar as the two agents are at the same layer of the hierarchy in the centralized world, and at different layers in the decentralized one. However, this seems to be based on the notion that the structure of unobserved side-contracting (which is hidden, illegal or informal) reflects the structure of the regular or permitted contracts in the respective hierarchies. Their theory provides no account of why this may be so.

The Laffont-Martimort (1998) model focuses on the restrictive case of the standard model with two agents, i.e., where costs take two possible values for each agent, and tasks are perfectly complementary. In centralization with collusion, given the contract $\{t_i(\theta_1, \theta_2), a_i(\theta_1, \theta_2)\}_{i=1,2}$ designed by P, the two agents enter into an *ex ante* side contract which decides on a coordinated set of cost reports $(\hat{\theta}_1, \hat{\theta}_2)$ submitted to P, and hidden side-transfers (s_1, s_2) as a function of cost reports they exchange privately after they observe their own true costs. This side contract is chosen to maximize the sum of their *ex ante* payoffs

$$(6) \quad E\left[\sum_{i=1}^2 [s_i(\theta_1, \theta_2) + t_i((\hat{\theta}_1, \hat{\theta}_2)(\theta_1, \theta_2)) - \theta_i q_i((\hat{\theta}_1, \hat{\theta}_2)(\theta_1, \theta_2))]\right],$$

subject to incentive and participation constraints within the coalition, besides the budget balance condition on the side-payments within the coalition. The incentive constraint states that each agent has an incentive to report his cost truthfully within the coalition: $\tilde{\theta}_i = \theta_i$ maximizes

$$(7) \quad E_{\theta_j} [s_i(\tilde{\theta}_i, \theta_j) + t_i((\hat{\theta}_1, \hat{\theta}_2)(\tilde{\theta}_i, \theta_j)) - \theta_i q_i((\hat{\theta}_1, \hat{\theta}_2)(\tilde{\theta}_i, \theta_j))],$$

while the participation constraint states that each agent must attain at least the expected utility he would attain by not entering the collusive agreement and playing the mechanism designed by P noncooperatively:

$$(8) \quad E_{\theta_j} [s_i(\theta_i, \theta_j) + t_i((\hat{\theta}_1, \hat{\theta}_2)(\theta_i, \theta_j)) - \theta_i q_i((\hat{\theta}_1, \hat{\theta}_2)(\theta_i, \theta_j))] \geq E_{\theta_j} [t_i(\theta_i, \theta_j) - \theta_i q_i(\theta_i, \theta_j)]$$

for each possible realization of θ_i .

In contrast in the delegation setting, Laffont and Martimort assume that A_1 makes a take-it-or-leave-it offer of an unobserved side contract, in which they again coordinate the cost report that A_1 makes to P on their joint behalf, and enter into side-transfers unobserved by P. There are two main differences in the nature of this side-contracting problem. First, the objective function is different: it is the expected utility of A_1 alone. Second, the participation constraint is different. Since P does not offer a formal contract to A_2 , and delegates this authority to A_1 , A_2 does not have the option of participating in

another mechanism should he refuse the side-contract offered by A_1 . The right-hand-side of (8) is thus zero for A_2 in the case of delegation.

Using the restrictive version of the standard model with *ex ante* symmetric agents, Laffont and Martimort (1998) find that delegation and centralization both achieve equivalent (second-best) profits for P, irrespective of whether or not there is collusion. This is no longer the case when there is an additional constraint of symmetric treatment of the agents — which they interpret either as a ‘fairness’ requirement, or the result of communication or information processing costs wherein P can only keep track of the aggregate cost $\hat{\theta}_1 + \hat{\theta}_2$ reported by the two agents, rather than their separate cost reports. They find then that delegation still achieves second-best profit, but centralization does not. The control of collusion requires asymmetric treatment of the two agents, which is not permitted in the centralized regime, but is possible in the delegation setting owing to the asymmetry in bargaining power that it allows through the side contract.

Bargaining power shifts to A_1 in the delegation setting for two reasons in the Laffont-Martimort (1998) theory. The first results from the assumption that the implied welfare weight on A_2 ’s expected payoff is zero under delegation, but equal to that on A_1 ’s payoff under centralization. This is based on an implicit assumption of a change in the bargaining protocol itself as a result of a shift in the formal contracting regime. The second reason for a diminution of A_2 ’s relative bargaining power is that he no longer has a backup option of participating noncooperatively.

The second reason for reduction in bargaining power of A_2 is intrinsic to the nature of the contracting: delegation gives no backup noncooperative option to A_2 should he refuse to collude, unlike centralization. The first reason is less intrinsic: it seems conceivable that the bargaining protocol itself does not change as a result of the formal contracting regime. For instance if the two agents belong to a union or a cartel or an industry association with ‘fairness’ norms which require equal treatment, and this association mediates the collusion, the allocation of welfare weights would not change across the two regimes. It therefore seems more natural to keep the bargaining protocol itself unchanged by the contracting regime, and focus only on the second reason for alteration in bargaining power which is entirely intrinsic to the regimes concerned.

For this reason, most of the subsequent literature assumes that welfare weights do not change between centralization and delegation; only the participation constraint for the side contract differs. It is assumed that in both regimes, the same agent A_1 makes a take-it-or-leave-it offer of side contract to A_2 .

If collusion is modelled in this way, it is evident that delegation cannot outperform centralization. A version of the Revelation Principle reappears: any outcome achieved by delegation can be replicated in centralization with P designing a contract which is null for A_2 , i.e, offers no transfers or production assignments to this agent. Then the nature of the constraints in P’s contracting problem are exactly the same, since the induced side contracting games are identical: the same objective function (A_1 ’s payoff) is maximized by the coalition, and the incentive and participation constraints are unchanged.

The question to be resolved, then, is whether delegation is costly relative to centralization. In other words, is it valuable for P to contract with both agents rather than

just one of them? Doing so allows her to manipulate the bargaining power within the coalition.

One additional ingredient of the models need to be mentioned: the treatment of participation constraints in the contract offered by P. Some authors (such as Mookherjee and Tsumagari (2004)) assume these are *ex post*, i.e., the agents decide whether or not to agree to participate after exchanging cost reports within the coalition. Others such as Faure-Grimaud, Laffont and Martimort (2003) and Celik (2009) assume these are *interim*: participation decisions must be made at the outset before the agents have had an opportunity to communicate with each other.

Mookherjee and Tsumagari consider the standard model with continuous types with two agents; they compare centralized contracting ($P^* - (A_1, A_2)$) with delegation to one of the productive agents ($P - A_1^* - A_2$), and to delegation to an intermediary ($P - M^* - (A_1, A_2)$) who is perfectly informed about the agents' costs.⁹ Using *ex post* participation constraints (in the sense described above where agents can collude in their participation decisions as well as in reports), they find the same rankings between these regimes hold as in the case of no collusion. Specifically, centralized contracting is always superior to delegation to one of the agents, while the ranking of the former *vis-a-vis* delegation to M depends on whether the two agents supply substitute or complementary inputs. Similar results also obtain in the discrete types case where M is not perfectly informed about the agents' costs, provided M is 'sufficiently' well-informed. These results provide an explanation of Chandler's observation that authority tends to be predominantly delegated by owners to managers who are pure intermediaries and not personally involved in productive tasks; their principal responsibility is to supervise and coordinate complementary tasks carried out by different workers.

Faure-Grimaud, Laffont and Martimort (2003) and Celik (2009) consider analogous questions in a setting with one agent A and one supervisor M , with a discrete set of cost types and where M is better-informed than P but not perfectly informed about A 's cost. While they both assume participation decisions in the mechanism offered by P must be made at the interim stage, their models differ with respect to details of the information structure. Faure-Grimaud, Laffont and Martimort study the two cost type case, where M receives a signal which also takes two possible values. They find that delegation to M (i.e., $(P - M^* - A)$) and centralization ($P^* - (M, A)$) achieve equivalent payoffs for P. Celik considers a case with three cost types, where M 's information consists of a partition of the state space. He finds that delegation to M does strictly worse. Hence the cost of delegation in these papers seems very sensitive to fine details of the information structure.

The results of Mookherjee and Tsumagari (2004) and Celik (2009) are similar and have a similar intuition. Collusion renders inoperable the second-best optimality of delegation, owing to the DMR problem, i.e., the adverse selection problem between the manager and his subordinate. The incidence of this problem can be reduced by raising the bargaining

⁹Baliga and Sjostrom (1998) compare centralized contracting with delegation to one of the agents in a setting with moral hazard with limited liability, rather than adverse selection. They find that the two are equivalent in a wide variety of circumstances, though not in all.

power of the subordinate *vis-a-vis* the manager. This can be achieved by P contracting with both agents, so as to allow A_2 (or A in the case of the Celik model) a positive outside option if he refuses the side contract offered by A_1 (or M in the case of the Celik model). However it is not easy to provide a simple intuition for why the same does not happen in the case considered by Faure-Grimaud, Laffont and Martimort (2003).

Motta (2009) argues that the analyses of Faure-Grimaud, Laffont and Martimort (2003) and Celik (2009) impose the overly restrictive assumption that contracts offered by P in the centralized regime are always accepted by both M and A on the equilibrium path. He shows that if we continue to assume that both parties must decide on whether or not to participate in P's mechanism at the interim stage, i.e., before they can collude, then there exists a centralized mechanism in which the second-best outcome can be achieved. This is true for either of the type or information structures considered by Faure-Grimaud, Laffont and Martimort (2003) and Celik (2009). The mechanism constructed by Motta involves offering a menu of contracts to A in which some options selected by A are associated with M not being employed to supervise A . These are tantamount to allowing an 'amnesty' scheme or 'self-reporting' by A which obviate the necessity of supervision. Motta shows that such schemes are useful in combating collusion, irrespective of how bargaining power is distributed within the coalition. Hence using such mechanisms which further augment the bargaining power of A allows the problem of collusion to be completely overcome in the centralized regime. This is irrespective of the structure of information available to the supervisor. At the same time, delegation continues to be unable to achieve the second-best, owing to the weak bargaining power of A in that regime. This suggests that centralization is superior to delegation in the presence of collusion, when agents cannot collude in their participation. Whether this continues to be so for the more relevant case where agents and supervisors can collude in their participation decisions as well, has not yet been studied.

In summary, the dust is yet to settle on how to model collusion and what it implies for the optimality of delegation. If delegation and centralization do not change bargaining weights of colluding participants, delegation cannot be superior to centralization. Whether it is inferior to centralization depends on the context and fine details of how collusion is modeled.

4. CONTRACTUAL INCOMPLETENESS, RENEGOTIATION AND COMPLEXITY

The preceding section considered contexts of 'complete' contracts, where the Principal can commit to a comprehensive mechanism in which there are no restrictions on the ability of agents to communicate their private information, of the Principal to process this information, make decisions and communicate corresponding instructions back to agents. These assumptions, combined with noncooperative behavior of agents, allows the Revelation Principle to hold, implying that delegation cannot out-perform centralization. To understand the benefits of delegation, it is therefore necessary to explore settings where the assumptions underlying the Revelation Principle fail to apply. Since we saw in the last section that collusion among agents does not provide a convincing theory of the benefits of delegation, it is necessary to explore some of the other assumptions. This

confirms the well-known proposition (going back to Williamson (1975) and reiterated by Hart (1995)) that one needs an ‘incomplete contract’ framework to have a cogent theory of allocation of control rights within an organization.

Contractual ‘incompleteness’ is however a very broad notion and there are different approaches to modeling it. The simplest approach is to exogenously rule out complete contracts, and impose ‘realistic’ restrictions on contracts — e.g., limiting their duration, or the kinds of contingencies they can incorporate. Aghion and Tirole (1997), Dessein (2002) and Bester and Kraemer (2008) use this approach to model the costs and benefits of delegation. Since these are included in the chapter on authority by Bolton and Dewatripont in this volume, we do not include a detailed discussion of these papers. We focus instead on theories where ‘incompleteness’ arises endogenously from either of two additional frictions: inability of the Principal to commit, and costs of communication or contractual complexity.

4.1. Commitment Problems. In many contexts it can be difficult for P to commit to deviate from *ex ante* promises to behave in a certain way in particular *ex post* contingencies, even though the latter can be anticipated and specified in advance. The reason is that *ex ante* decisions promised by P may not be optimal *ex post*, as the former may have been prompted by the need to provide certain kinds of incentives to A. Once the required actions have been chosen by A, the P may wish to revert to a different decision which is in the subsequent *ex post* mutual interest of P and A. Delegation of authority to an ‘independent’ third party may be a way to overcome this commitment problem.

Examples of this ‘time-inconsistency’ problem in the context of monetary policy have been familiar to macroeconomists since the work of Kydland and Prescott (1977). Governments with a short-term interest in expansionary monetary policy tend to create excessive inflation when they possess discretionary control. This forms the primary motivation for delegation of monetary policy to an independent Central Bank (Rogoff (1985)). The solution to the problem of commitment to an anti-inflationary monetary policy is for the government to delegate day-to-day policy-making to bureaucrats or technocrats with a known commitment to anti-inflationary policy, and insulate them from short-term political pressure. Similar problems arise also in fiscal policy, such as the problem of bailouts of public sector firms or ‘soft budget constraints’ (see, e.g., Dewatripont and Maskin (1995), Qian and Weingast (1997)), or in income tax audits (Melumad and Mookherjee (1989)). Cremer (1995) provides an interesting model of commitment problems in organizations along similar lines, in the context of design of employment contracts: owners will find it difficult to follow through on threats to fire high quality agents detected shirking, if these agents are expected to be superior relative to others who might replace them. Delegation of employment decisions to a manager who is provided strong incentives to fire unproductive workers via an incentive scheme, may be a way of committing to implement termination threats which induce workers not to shirk. One problem with this explanation of delegation is that it is based on an implicit assumption that it is easier for Principals to commit to delegating the allocation of control rights, than it is to commit to detailed decisions if they retain the rights themselves. There may be problems arising even with these delegation arrangements, wherein the Principal may

seek to reallocate control rights *ex post*, or attempt to influence the decisions made by agents to whom authority has been delegated (see Katz (1991) for a detailed analysis). In Cremer's model for instance, if employment decisions were delegated to a manager who is supposed to be penalized for retaining unproductive workers, the owner would have an *ex post* incentive for not applying this penalty when the worker in question was discovered to be a high quality type. Neither the worker or manager would have an interest in opposing P's opportunistic intervention. Hence there could be problems with the enforcement of commitments to delegate. It is not clear whether and why it is easier to commit to delegate authority than to actual decisions.

A tighter argument for the value of delegation on the basis of renegotiation problems is provided by Beaudry and Poitevin (1995) and Poitevin (1995, 2000) in the context of a standard principal-agent production problem. Opportunities for (*ex post* Pareto-improving) renegotiation are provided in both centralized and decentralized environments. They show that the decentralized regime is less vulnerable to renegotiation, and can accordingly generate a higher level of payoff for the Principal.

The argument is simplest in a single agent setting. Suppose the agent A produces a at a cost of $\theta \cdot a$, which generates revenue $B(a)$ to P, and A is privately informed about the realization of θ . Under centralization the contracting game is as follows. In Stage 1, P offers a contract $a(\theta), t(\theta)$. At Stage 2, A observes the realization of θ , decides whether to participate and submits a cost report to P. At Stage 3, P can offer a new contract to A. If A accepts, the game continues with the new contract in force. Otherwise the previous contract continues to apply. The output and transfer are then determined by the prevailing contract, and the report already submitted by A.

Attention can be restricted to contracts which are not renegotiated on the equilibrium path. It can then be shown that the equilibrium renegotiation-proof allocation must be separating, i.e., P will learn the true cost of A at the end of Stage 2.¹⁰ Hence P will subsequently offer an *ex post* efficient allocation at Stage 3. This implies that the second-best solution, i.e., the allocation associated with a commitment contract, cannot be supported in the presence of opportunities for renegotiation, as long as it involves *ex post* distortions.

Now consider delegation, wherein P offers A an incentive contract $t(a)$ at Stage 1. At Stage 2, A observes θ , decides whether to participate and selects a . At Stage 3, P observes a and can offer a new contract. But given that a has already been chosen, the only scope for renegotiation is over the transfer payment. But there cannot be any scope for an *ex post* Pareto-improving renegotiation over the transfer payment alone. Hence the second-best allocation can be implemented under delegation. It follows that delegation typically dominates centralization as it is less vulnerable to renegotiation. The key difference is that in centralization, decisions over production *and* transfers are made by P following communication from the agent. In delegation renegotiation can happen after the production decision has already been made by A.

Does this suggest that delegation generally out-performs centralization? Poitevin (1995) points out this is not the case when P also receives relevant information after signing

¹⁰See Beaudry-Poitevin (1995, Proposition 5) for details.

the initial contract. Suppose P learns the value of a ‘market’ parameter η that enters the revenue function: $B(a; \eta)$ at the same time that A learns the realization of cost θ . Delegating production decisions to A without any communication from P regarding the realization of η would then imply independence of production decisions with regard to the realization of market information. Conversely, the benefit of centralization based on communication of cost information from A would be that production decisions would be responsive to market information, at the cost of being vulnerable to renegotiation with A. There are now four possible organizational alternatives: pure centralization (decisions are made by P without communication from A), pure decentralization without communication (production decisions are made by A without communication from P about the realization of η), and corresponding versions with communication (which Poitevin calls a *hierarchy*). A hierarchy is more responsive to information about both cost and market parameters, at the cost of being more vulnerable to renegotiation.

4.2. Communication and Complexity Costs. Melumad, Mookherjee and Reichelstein (1992, 1997) explore the implications of costly communication or contractual complexity for the costs and benefits of delegation *vis-a-vis* centralization, in the context of the ‘standard’ production model with one Principal and two agents. Their 1992 paper focuses on communication costs, which are argued to arise from expertise of agents not shared by the Principal, which makes it impossible for the agents to communicate all they know to the latter. Languages used by Principal and agents may not be exactly the same, and they could lack a rich enough shared vocabulary. There could also be limits on the Principal’s ability to process the reports sent by the agents, within a given time limit within which decisions need to be made.

Such restrictions are incorporated as an (exogenous) finite limit on the message set available to any agent for communicating with P (under either centralization or decentralization), while the agent’s private information is real-valued. This implies that full ‘revelation’ is impossible: agents must compress what they know into messages in a way that entails loss of information. The finite limit could represent a common finite alphabet used for communication. Alternatively if sending each ‘bit’ of information takes one unit of time, then a given time limit for making decisions imposes an upper bound to the number of ‘bits’ that can be communicated. This is similar to the approach used in computer science to measure communicational complexity, e.g., i.e., ‘tree complexity’ of communication protocols: Karchmer (1989), Segal (1995). Nevertheless, Melumad, Mookherjee and Reichelstein do not model the underlying sources of the communicational restrictions. Instead they impose a common restriction on the size of message sets which applies uniformly to both centralization and decentralization, and compare their performance under any given size restriction. This follows in the tradition of earlier models of Green and Laffont (1986, 1987). A similar approach was adopted by Laffont and Martimort (1998) in which the Principal was assumed to only be able to receive aggregates of cost reports sent by different agents.

The Melumad-Mookherjee-Reichelstein (1992) model imposes a number of additional restrictions on communication and contracting. Their version of centralization in the standard two-agent setting is the following. Agent i ’s message set is restricted to contain

$k_i (< \infty)$ elements, whereas his type space is a one-dimensional interval. At Stage 1, P offers a contract $\{a_i(m_1, m_2), t_i(m_1, m_2)\}_{i=1,2}$ defined over message-tuples $(m_1, m_2) \in M_1 \times M_2$, where M_i contains k_i elements. At Stage 2, each agent observes his cost θ_i , decides whether to participate and reports m_i . The agents either report simultaneously, or sequentially. These reports determine production assignments and transfers as per the stipulated contract.

An important restriction here is that there is only one round of communication per agent. In a revelation mechanism where there is no restriction on the message set size, this is without loss of generality since one round of communication suffices to communicate the entire information possessed privately by the agent. But when message set sizes are restricted, one round of communication will not suffice for the agent to communicate everything he knows, and there is a rationale for multiple rounds of communication. This is ruled out by assumption, in order to keep the analysis simple. We shall return to this issue later in this section.

Under delegation, the game proceeds as follows. At Stage 1, P offers A_1 a contract $t_1(B, t_2; m_1)$. At Stage 2, A_1 observes θ_1 , decides whether to participate, reports m_1 to P , and offers A_2 a subcontract $t_2(m_2; m_1, \theta_1), a_2(m_2; m_1, \theta_1)$. At the same time A_1 makes a plan to produce $a_1(m_2; m_1, \theta_1)$. At Stage 3, A_2 observes θ_2 , decides whether to participate, and reports m_2 to A_1 . Note here that while message sets are restricted, no analogous restriction is imposed on how finely the subcontract offer can vary with θ_1 , Agent 1's true type. Implicit here is that there are no restrictions on the ability of A_1 to communicate details of the sub-contract to A_2 .

The main result is that *for any finite message set restrictions k_1, k_2 , delegation outperforms centralization*. A similar result is obtained in their follow-up 1997 paper, in which restrictions on communication are derived from restrictions on complexity of contracts, where the latter is measured by the total number of contingencies (which correspond to message combinations in centralization, and message-action combinations in decentralization).

The underlying idea is that finite communicational restrictions prevent agents from fully revealing their real-valued cost realizations to the Principal under centralization. Delegating decision-making authority to agents would enable decisions to be based on greater 'local' information, resulting in a *flexibility* gain. Offsetting this is the possibility that this authority can be 'abused' by agents to their own advantage, resulting in a *loss of control*: e.g., double marginalization of rents. The ability of the Principal to moderate this control loss depends upon the richness of communication received from agents and her ability to monitor their performance, which are also restricted by the nature of the contracts. The tradeoff between centralized and decentralized decision-making is then based on a comparison of the flexibility gain and the control loss. Under the sequencing, monitoring and risk neutrality assumptions described in Section 3, it turns out that delegation strictly dominates centralization.

At the same time, if P lacks the ability to monitor detailed production assignments or transfers between the agents, examples can be constructed where the control losses inherent in delegation may outweigh the gains in flexibility to ensure that centralization

is superior. Hence the theory provides contexts where either one of the two modes is superior, depending on the ability of P to monitor the side contract.

These results thus provide a rationale for studying conditions for delegation to be optimal with ‘complete’ contracts: the very conditions that ensure this turn out to ensure that delegation is superior to centralization when communication or contractual complexity is costly. When delegation is inferior with complete contracts, it may also be so when contracts are incomplete owing to communication or complexity costs. In general, there is a trade-off between the ‘flexibility’ advantage of delegation (which arises only when contracts are incomplete) and its control loss (which arises also when contracts are complete).

Nevertheless a number of problems remain with this theory, as discussed in a recent paper by Mookherjee and Tsumagari (2008). For one, the restriction to a single round of communication is *ad hoc*: most real world organizations are characterized by interactive multistage communication between agents. The mechanism design problem includes the design of communication protocols, a challenging problem given finite restrictions on the size of message sets and the number of rounds of communication. Mookherjee and Tsumagari consider a context where agents take time to read and write messages, and production decisions need to be made within a finite time horizon, combined with time taken to read and write messages. This allows a wide range of finite communication protocols, varying with regard to the number of rounds and message set size at each round.

Another restrictive feature of the Melumad-Mookherjee-Reichelstein (MMR) theory is that it focuses only on polar extremes of a wider range of possible mechanisms. A mechanism can be thought of as including three classes of decisions: contracting, communication and production, each of which can be centralized or decentralized. The centralized mechanisms in MMR centralize all three sets of decisions: P contracts with all agents; each agent communicates only with P , and P makes all decisions concerning transfers. The delegation mechanism distributes responsibility for subcontracting and communication with A_2 to A_1 . Intermediate mechanisms could also be considered, e.g., where contracting is centralized, but communication and production decisions are decentralized (as in the prototypical large ‘Japanese’ organization where central headquarters employs all workers, but substantial responsibility for discussing and resolving shop-floor problems are delegated to worker teams: see Aoki (1990)). In the simple setting of the two agent production model, P could design an incentive contract for each agent in which the bonus for producing a larger number of units of their respective outputs depends on messages exchanged between them. Coordination can be achieved by conditioning the bonuses on cost messages of all agents, while flexibility is achieved by decentralizing production decision to respective workers.

Mookherjee and Tsumagari develop a theory of mechanism design which allows a broad class of finite communication protocols with multiple stages, as well as a wide range of mechanisms where different components of contracting, communication and production can be centralized or decentralized. Based on an assumption that messages exchanged between agents are costlessly verifiable by P , they show that centralized contracting can generally perform as well as decentralized contracting. At the same time, decentralized

production and communication systems generally outperform centralized ones — owing to the greater flexibility they allow in choice of production and coordination of information across agents. In other words, it indicates the optimality of the prototypical ‘Japanese’ organization.¹¹ Hence even within the simple context of a two-layer hierarchy the range of mechanisms that ought to be considered is wider than the canonical choice between the polar ‘all-centralized’ and ‘all-decentralized’ organizational forms.

5. MORE COMPLEX HIERARCHIES

The preceding two sections have been concerned with foundational issues in the modeling of costs and benefits of delegating control to managers or intermediaries. Accordingly they focused on simple contexts involving one principal or owner, one manager and one or two agents who play a productive role. In this section we describe models of more complex hierarchies, involving more agents or departments, which help address a number of additional questions concerning the organization of hierarchies. An example of such a question concerns grouping of activities into different ‘departments’. In *Strategy and Structure*, Chandler described the evolution of many large vertically integrated firms producing a multitude of products in various regions or markets from unitary (U-form) to multidivisional (M-form) organizations. In the former, departments are organized along functional lines: departments concentrate plants and activities producing the same input or intermediate good. In the latter, they are organized by product or market or region, with each division being internally vertically integrated and relatively independent of other divisions (except for overhead, capital or top management allocated by top management across divisions). Socialist economies likewise can be organized by ministries producing specific intermediate inputs (as in the erstwhile Soviet economy), or by regions each of which is relatively self-contained (as in the Chinese economy). Chandler argued that the shift away from the U-form freed up the time and effort of central management from day-to-day management of operations in each division that was required owing to the need to coordinate activities of different producing divisions. In the M-form they could delegate internal management of divisions to divisional managers, and focus more on long-range issues. However this meant giving up on the advantages of scale economies and specialization possible from the concentration of similar activities within the same division in the U-form.

Understanding this dimension of organizational design requires extension of the models described in previous section to incorporate multiple divisions and problems of horizontal coordination across these divisions. The previous models can be thought of as organizations consisting of a single division, with focus on problems of vertical control and coordination.

Rotemberg (1999) provides an interesting model in which the key trade-off between an U-form and a M-form hierarchy involves coordination versus control. A firm produces two products A, B and two associated intermediate inputs 1, 2 each of which is required in the production of either good. There are four agents or plant managers, represented by

¹¹When however communication between agents is difficult to verify by employers, they provide an example where decentralized contracting outperforms centralized contracts.

a combination of a product and an input used for that product. Rotemberg assumes that a pair of agents will have a common supervisor, and thereby form a division. Implicit here is an exogenous upper bound on the span of control: a supervisor cannot oversee more than two agents. This rules out a fully centralized firm with a single division which contains all four agents. Rotemberg assumes therefore the firm has two divisions and two corresponding supervisors, who jointly own it. Another important assumption is that the supervisors operate independently and cannot share any information with one another.

In the U-form, divisions are organized by the inputs produced: division i produces input i and includes the two plants iA, iB producing this input, where $i = 1, 2$. In the M-form, they are organized by products: division j produces product j and includes the plants $1j, 2j$ producing the two inputs needed for this product, where $j = A, B$. Each plant ij has to choose a level of activity x_j^i and a method m_j^i . The inputs used in the production of each product need to be coordinated: there is a loss function for each product j which depends on the difference in activity levels x_j^1, x_j^2 . The M-form allows better coordination since it groups different plants involved in the same product group into the same department.

The optimal method of producing the same input across different products is the same. Incentive problems arise with respect to choice of method, an issue regarding the costs of which the respective plant managers are privately informed. The main advantage of the U-form is that the supervisor controls two agents producing the same input, thus allowing the report of each agent to discipline the other. Hence the U-form allows for better control of incentive problems.

The model has a number of interesting predictions. The U-form involves less high powered incentives for plant managers, owing to the stronger information available to the corresponding supervisor. As the firm expands in scale, it adds more plants within each division. In the M-form there are now multiple plants producing the same input, which helps reduce the incentive problem. But the U-form continues to be subject to coordination losses. Hence for sufficiently large scale, the M-form must dominate. This helps explain how the American firms shifted from one form to the other as described by Chandler: the coordination problems with the U-form got worse relative to the control problems as firms grew in scale.

Mookherjee and Reichelstein (1997, 2001) provide a model of hierarchies which generalizes the standard production model in preceding sections to the context of many agents, departments and products. They provide general sufficient conditions on the structure of the organizational production function or technology for a hierarchy with an arbitrary number of branches and layers to be able to achieve second-best profits for the Principal. The mechanism they use extends the ‘responsibility center’ mechanism described in Section 3 for the case of two agents that form a single division. With multiple divisions that constitute different branches of the hierarchy at any given layer, the mechanism has to ensure horizontal coordination across these branches as well as vertical coordination and control across layers within any given branch. In addition to the assumptions concerning sequencing of contracts, absence of collusion and of risk-aversion or limited liability

constraints, the key assumption required is that the organization of the hierarchy is *consistent* with the technology. This means that the technology is recursively decomposable into the production of intermediate inputs at different stages of production, and each plant has constant returns to scale.

Absent incentive problems, this consistency condition implies that an optimal production plan can be formed for the firm as a whole, with (one-dimensional) cost reports (or budgets or forecasts) sent by agents that flow up the hierarchy, which are aggregated at each layer and passed up by each manager to his manager in turn. Subsequently production targets are formed at the top of the hierarchy and these then flow down the hierarchy. Each divisional manager forms a contingent production plan, allocating production responsibility across subordinate divisions on the basis of their respective cost reports, and the aggregate production target assigned to the division. Mookherjee and Reichelstein show that the same kind of mechanism can be used even in the presence of incentive problems, where each agent is self-interested. It requires an initial phase is added wherein contracts flow down the hierarchy, in which the manager at any layer offers a contract to each subordinate, which the latter must respond to prior to offering contracts to their subordinates in turn. The assumptions required are the same as those that ensure optimality of delegation in a three-layer hierarchy: top-down contracting, monitoring of ‘costs’, and risk-neutrality of all agents. The construction resembles the Calvo-Wellisz method of ‘replicating’ the three-layer hierarchy to accommodate more layers and branches. Under these conditions, then, there is no limit to the size of the hierarchy.

Applied to the U-form versus M-form question, this model provides conditions when either can be optimal. The U-form is optimal if it is consistent with the technology — as would be the case if the firm produces a single product by combining different inputs each of which is produced by multiple plants. Each division then corresponds to the production of a particular input. The M-form is consistent with the technology if the firm produces multiple products, each of which combines a number of inputs (some of which may be used in different products). The divisions are then organized along product lines. Each division is independent of others, except for allocation of some overhead facilities or capital resources by headquarters. When the technology takes this form, it is evident that the U-form is not consistent with it. This approach thus suggests the role of changing technology and product variety in explaining evolution from U-form to M-form organizations. With the emergence of multiple products produced by the same firm, the U-form would experience problems relative to the M-form in coordinating production of different inputs used in any given good (in the sense of being able to use the ‘budgeting’ system with one-dimensional cost and quantity aggregates in every division).

An alternative perspective on U-form versus M-form organizations is provided by Maskin, Qian and Xu (2000). Their theory abstracts from coordination issues altogether, and focuses on informational and incentive implications of these alternatives. Similar to the basic Rotemberg model, the Maskin-Qian-Xu model has a three layer hierarchy producing two products $i = 1, 2$ in two different regions $r = A, B$. There are four plants, corresponding to a combination of a product and a region. Productivity shocks arise at the economy, regional and industrial levels. The organizational structure assigns

responsibility to managers to devote costly effort to improving productivity. Observed performance levels are the confluence of effort and shocks, giving rise to moral hazard problems. In both U-form and M-form a top manager deals with the economy-wide shock. The essential difference arises at the middle layer of the hierarchy. In the U-form, departments are organized by industry, so there is a manager at the middle layer in each department whose effort affects the productivity of the two plants in the corresponding industry. In the M-form the departments are organized by region, so the middle-level manager's effort affects the performance of the two plants in two different industries which happen to be located in the same region. Bottom-level managers' efforts affect the plants they are assigned to. In the U-form (resp. M-form) these correspond to different regions (resp. industries) within the same industry (resp. region). The two organizational forms thus differ in the assignment of tasks across managers. In the M-form, for instance, only one manager (located at the middle tier) in the organization is dealing with the regional shocks, whereas there are two managers in the U-form (located at the bottom layer in two different industry divisions) who deal with regional shocks.

Managers are risk averse, giving rise to a trade-off between risk-sharing and effort incentives in the design of their compensation schemes. 'Yardstick' competition is an important way of dealing with these incentive problems, wherein agents performance is measured relative to that of peers whose performance is subject to similar (i.e., correlated) shocks. Maskin, Qian and Xu show that the U-form and M-form offer identical opportunities for evaluating performance of managers at the top and bottom tier, but they differ with regard to managers at the middle tier. If performance at the regional level (aggregating across industries) is more comparable (i.e., correlated) across regions, rather than performance at the industrial level (aggregating across regions), then the M-form allows more effective use of yardstick competition for middle-level managers. For top-layer managers, there is no opportunity for yardstick competition (as in the model this is the only firm in the economy). And for bottom level managers, there exists one other manager in the firm whose performance provides a comparable yardstick in both organizations. The model thus predicts that performance differences will arise owing to differential managerial effort incentives at intermediate levels of the hierarchy. Maskin, Qian and Xu go on to provide empirical evidence from Chinese firms that the structure of productivity shocks is such that performance is more comparable across regions than across industries. This suggests the superiority of the M-form, thus providing a potential explanation for superior performance of the Chinese over Soviet version of socialism.

6. CONCLUDING REMARKS

The literature overviewed in this chapter pertain mainly to the microfoundations of an incentive-based theory of hierarchies. The Holy Grail of this field has been a formalization of Williamson's ideas of control loss in hierarchies based on incentive problems, and how this affects the size and structure of organizations. This have proved difficult, mainly because addressing this question requires as a prior requirement an understanding of what is that managers 'do', and the costs and benefits of delegating to them authority. Only when there is a basic model of 'management' can the theory be used to address

detailed questions concerning the nature of control loss, and how it varies with the size and structure of the organization.

The models described here focus primarily on incentive problems associated with supervision and the more comprehensive notion of delegation of authority. They address questions of the design of compensation for supervisors, how these relate to compensation of production workers and attendant problems of collusion. The models help pose the question of choice amongst different organizational designs, highlighting the different dimensions involved in these designs even in the simplest contexts involving one or two productive agents and one manager: breadth (span of control) and depth (number of vertical layers); extent of responsibility delegated at each level and related compensation rules and monitoring systems; how contracts, communication and production planning are sequenced. So far most of the models have been rudimentary, with only one or two production workers, one manager and one owner. At the next step one expects these will be used as building blocks for more complex organizations and design issues in the presence of more agents, managers, products and intermediate goods.

Despite their simplicity, the models enable the age-old question of centralization versus decentralization to be posed in different concrete ways. They have also generated a wide range of applications in fields as diverse as management accounting (profit centers and budgeting¹², transfer pricing¹³, auditing¹⁴), fiscal decentralization (hard versus soft budget constraints¹⁵, treatment of externalities¹⁶, or accountability in local governments¹⁷), procurement and regulation¹⁸, and comparisons between Soviet-style and Chinese-style socialism¹⁹.

Nevertheless, much remains to be done. There is considerable scope and need to use these models to address more applied questions in industrial organization, and enable closer integration with empirical work on internal organization of firms. What are the effects of changes in information technology, competition on the product market or openness to trade on the internal structure of firms? These issues have been discussed extensively in fields of management (e.g., Hammer and Champy (1993)), and have been the subject of recent empirical studies (e.g., Acemoglu et al (2006), Brynjolfsson and Hitt (2000), Bresnahan, Brynjolfsson and Hitt (2002), Caroli and van Reenen (2001), Rajan and Wulf (2006)). With the exception of Acemoglu et al (2006), most of this literature lacks a theoretical framework. A closer integration of theory and empirics would enrich these analyses, and permit better understanding of normative implications.

Even within the context of purely theoretical analysis, significant challenges and open questions remain. An integration of incentive issues with costly information processing would represent a major step forward at the conceptual level. This would help formalize

¹²Mookherjee and Reichelstein (1997)

¹³Edlin and Reichelstein (1996), Vaysman (1996, 1998), Baldenius and Reichelstein (2004)

¹⁴Melumad and Mookherjee (1989), Kofman and Lawaree (1993)

¹⁵Dewatripont and Maskin (1995), Qian and Roland (1998)

¹⁶Klibanoff and Poitevin (2009)

¹⁷Seabright (1996), Bardhan and Mookherjee (2000, 2006))

¹⁸Baron and Besanko (1992), Gilbert and Riordan (1995), Severinov (2008).

¹⁹Jin, Qian and Weingast (2005), Maskin, Qian and Xu (2000), Qian and Weingast (1997)

hierarchy design as a trade-off between information processing advantages with incentive and control disadvantages of delegating authority. Issues concerning the design of more complex hierarchies still remain to be addressed, e.g., models of delegation which address questions of size and structure of hierarchies, and related implications for compensation, rents and performance at different levels. The simple models described in this chapter suggest the role of collusion, limited liability or managerial risk aversion as sources of control losses in hierarchies. They need to be embedded in more complex settings to develop detailed predictions concerning size and structure of firms and how they are affected by technology and market parameters. This would then help realize the goal of constructing a Williamsonian model of hierarchies with secure microfoundations, which can be empirically tested. There are no papers which have tried to formalize Williamson's ideas based on noise and errors in communication. Other important questions include comparisons of hierarchies with non-hierarchical organizational structures (such as pol-yarchies and matrix organizations) and interaction between formal and informal control systems (Baker, Gibbons and Murphy (1999)).

REFERENCES

- Acemoglu D. , P. Aghion, L Lelarge, J Van Reenen and F. Zilibotti (2006), "Technology, Information and the Decentralization of the Firm," NBER Working Paper No. 12206.
- Aghion P. and J. Tirole (1997) "Formal and Real Authority in Organizations," *Journal of Political Economy* 105(1), 1-29.
- Aoki M. (1990), "Toward and Economic Model of the Japanese Firm," *Journal of Economic Literature*, 28(1), March, 1-27.
- Baker G., R. Gibbons and K. Murphy (1999), "Informal Authority in Organizations," *Journal of Law, Economics and Organization*, 15, 56-73.
- Baliga, S. and T. Sjostrom (1998), "Decentralization and Collusion," *Journal of Economic Theory*, 83:196-232.
- Baldenius T and S. Reichelstein (2004), "External and Internal Pricing in Multidivisional Firms," Working Paper 1825(R), Graduate School of Business, Stanford University.
- Bardhan P. and D. Mookherjee (2000), "Capture and Governance at National and Local Levels," *American Economic Review*, Papers and Proceedings, 135-139.
- (2006), "Corruption and Decentralization of Infrastructure Delivery in Developing Countries," *Economic Journal*, 116, Jan 2006, 107-133.
- Baron, D. P. and D. Besanko (1992), "Information, Control, and Organizational Structure," *Journal of Economics and Management Strategy*, 1:237-275.
- Beaudry P. and M. Poitevin (1995) "Contract Renegotiation: A Simple Framework and Implications for Organization Theory," *Canadian Journal of Economics*, 28, 302-335.

- Bester H. and D. Krahmer, “Delegation and Incentives,” *RAND Journal of Economics*, 39(3), 664-682.
- Bolton, P. and M. Dewatripoint (1994) “The Firm as a Communication Network”, *Quarterly Journal of Economics*, 109, 4, 809-839.
- Bresnahan T., E. Brynjolfsson and L. Hitt (2002), “Information Technology, Workplace Organization, And the Demand for Skilled Labor: Firm-Level Evidence,” *Quarterly Journal of Economics*, February 2002, 339–376.
- Brynjolfsson, E. and L. Hitt (2000), “Beyond Computation: Information Technology, Organizational Transformation and Business Performance,” *Journal of Economic Perspectives*, 14(4), Fall 2000, 23–48.
- Calvo, G. and S. Wellisz (1978) “Supervision, Loss of Control and the Optimal Size of the Firm” *Journal of Political Economy*, vol. 86, 943-952.
- Caroli E. and J Van Reenen (2001), “Skill-Biased Organizational Change,” *Quarterly Journal of Economics*, 116, 1148-92.
- Celik G. (2009), “Mechanism Design with Collusive Supervision,” *Journal of Economic Theory*.
- Chandler A. (1962), *Strategy and Structure*, Cambridge: MIT Press.
- (1977), *The Visible Hand*, Cambridge, MA: Harvard University Press.
- Cremer, J. “Arms Length Relationships,” *Quarterly Journal of Economics*, 110(2), 275-295.
- Cremer, J. and M. Riordan (1987) “On Governing Multilateral Transactions with Bilateral Contracts,” *Rand Journal of Economics*, vol. 18, no. 3, 436-451.
- Datta S. (1996), “On Control Losses in Hierarchies,” *Rationality and Society*, 8(4), 387-412.
- Da Rocha J.M. and M.A. de Frutos (1999), “A Note On the Optimal Structure of Production,” *Journal of Economic Theory*, 89, 234-246.
- Dessein, W. (2002), “Authority and Communication in Organizations”, *Review of Economic Studies*.
- Dewatripoint M. and E. Maskin (1995), “Credit and Efficiency in Centralized and Decentralized Economies,” *Review of Economic Studies*, 62(4), 541-555.
- Edlin A. and S. Reichelstein (1995), “Specific Investment under Negotiated Transfer Pricing: An Efficiency Result,” *The Accounting Review*, April 1995.
- Faure-Grimaud, A. and D. Martimort (2001), “Some Agency Costs of Intermediated Contracting,” *Economics Letters*, 71(1):75-82.
- (1999), “Political Stabilization by an Independent Bureaucracy”, mimeo, London School of Economics.

- Faure-Grimaud, A., J.J. Laffont and D. Martimort (2000), "A Theory of Supervision with Endogenous Transaction Costs," *Annals of Economics and Finance*, 1, 231-263.
- Faure-Grimaud, A., J-J. Laffont and D. Martimort (2003), "Collusion, Delegation and Supervision with Soft Information," *Review of Economic Studies*, 70, 253-280.
- Gilbert R. and M. Riordan (1995) "Regulating Complementary Products: A Comparative Institutional Analysis," *Rand Journal of Economics*, 26, 243-256.
- Green J. and J. Laffont, (1986), "Incentive Theory with Data Compression," in W. Heller, R. Starr and D. Starrett eds, *Essays in Honor of Kenneth Arrow*, vol 3, Cambridge Univ Press.
- (1987), "Limited Communication and Incentive Compatibility," in T. Groves, R. Radner and S. Reiter eds, *Information, Incentives and Economic Mechanisms*, University of Minnesota Press.
- Hammer M. and J. Champy (1993) *Reengineering the Corporation*. New York: Harper Collins.
- Hart, O. (1995) *Firms, Contracts, and Financial Structure*. Oxford and New York: Oxford University Press, Clarendon Press, 1995.
- Ichniowski C., K. Shaw and G. Prennushi (1997), "The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines," *American Economic Review*, 87(3), 291–313.
- Jin H., Y. Qian and B. Weingast (2005), "Regional Decentralization and Fiscal Incentives: Federalism, Chinese Style," *Journal of Public Economics*, 89(9-10), 1719-42.
- Karchmer M. (1989), *Communication Complexity and Circuit Depth*, Cambridge MA: MIT Press.
- Katz M. (1991), "Game-Playing Agents: Unobservable Contracts as Precommitment," *RAND Journal of Economics*, 22(3), Autumn.
- Klibanoff P. and M. Poitevin (2009), "A Theory of (De)Centralization," working paper, Kellogg School of Management, Northwestern University.
- Kofman F. and J. Lawaree (1993), "Collusion in Hierarchical Agency," *Econometrica*, 61(3), 629-656.
- Kydland F. and E. Prescott, "Rules Rather than Indiscretion: Time Consistency of Optimal Plans," *Journal of Political Economy*, 85, 473-492.
- Laffont J. and D. Martimort (1998) "Collusion and Delegation," *Rand Journal of Economics* 29(2), 280-305.
- Macho-Stadler I. and J.D. Perez Castrillo (1998), "Centralized and Decentralized Contracts in a Moral Hazard Environment," *Journal of Industrial Economics*, 46(4), 489-510.
- Maskin E., Y. Qian and C. Xu (2000), "Incentives, Information and Organizational Form," *Review of Economic Studies*, 67, 359-378.

- McAfee, P. and J. McMillan (1995) "Organizational Diseconomies of Scale" *Journal of Economics and Management Strategy*, 4(3), 399-426.
- Melumad N. and S. Reichelstein (1989), "Value of Communication in Agencies," *Journal of Economic Theory*, 47(2), 334-368.
- Melumad N. and D. Mookherjee (1989), "Delegation as Commitment: The Case of Income Tax Audits", *RAND Journal of Economics*, 20(2), 139-163.
- Melumad, N., D. Mookherjee and S. Reichelstein (1995) "Hierarchical Decentralization of Incentive Contracts" *Rand Journal of Economics*, 26(4), 654-672.
- (1992) "A Theory of Responsibility Centers" *Journal of Accounting and Economics* vol. 15, 445-484.
- (1997) "Contract Complexity, Incentives and the Value of Delegation," *Journal of Economics and Management Strategy* 6(2), 257-289.
- Mookherjee, D. and S. Reichelstein (1997), "Budgeting and Hierarchical Control," *Journal of Accounting Research*, 35(2): 129-55.
- Mookherjee, D. and S. Reichelstein (2001), "Incentives and Coordination in Hierarchies," *BE Journals Advances in Theoretical Economics*, 1(1).
- Mookherjee, D. and M. Tsumagari (2004), "The Organization of Supply Networks: Effects of Delegation and Intermediation," *Econometrica*, 72(4), 1179-1220.
- (2008), "Mechanism Design with Limited Communication: Implications for Decentralization," mimeo, Department of Economics, Boston University.
- Motta A. (2009), "Collusion and Selective Supervision," working paper n. 93, Universita Degli Studi Di Padova.
- Mount K. and S. Reiter (1974), "The Informational Size of Message Spaces," *Journal of Economic Theory* 8, 161-191.
- Mount K. and S. Reiter (1995), "A Theory of Computing with Human Agents," working paper, Department of Economics, Northwestern University.
- Myerson R. (1982), "Optimal Coordination Mechanisms in Generalized Principal Agent Problems," *Journal of Mathematical Economics*, 10, 67-81.
- Poitevin, M. (1995), "Contract Renegotiation and Organizational Design," mimeo, CIRANO Working Paper No. 95-3, Montreal.
- Poitevin, M. (2000), "Can the Theory of Incentives Explain Decentralization?," *Canadian Journal of Economics*, 33(4):878-906.
- Qian Y. (1994), "Incentives and Loss of Control in Optimal Hierarchy," *Review of Economic Studies*, 61(3), 527-544.
- and B. Weingast (1997), "Federalism as a Commitment to Preserving Market Incentives," *Journal of Economic Perspectives*, 11(4), 83-92.

- Radner, Roy (1992) "Hierarchy: The Economics of Managing" *Journal of Economic Literature* vol. 30, 1382-1415.
- (1993) "The Organization of Decentralized Information Processing," *Econometrica* vol. 61, no. 5, 1109-1146.
- Radner R. and J. Marschak (1972), *Economic Theory of Teams*, New Haven: Yale University Press.
- Rajan R. and J. Wulf (2006), "The Flattening Firm: Evidence From Panel Data on the Changing Nature of Corporate Hierarchies," *Review of Economics and Statistics*, 88(4), 759-773.
- Rogoff K (1985), The Optimal Degree of Commitment to an Intermediate Monetary Target, *Quarterly Journal of Economics*, 100 (November 1985), 1169-1189
- Rotemberg J. (1999), "Process Versus Function Based Hierarchies," *Journal of Economics and Management Strategy*, 8(4), 453-487.
- Seabright P. (1996), "Accountability and Decentralization in Government: an Incomplete Contracts Model", *European Economic Review*.
- Segal, I. (1995), "Communication Complexity and Coordination by Authority," *mimeo*, Department of Economics, University of California, Berkeley.
- Severinov, S. (2008), "The Value of Information and Optimal Organization " *RAND Journal of Economics*, 39[1], Spring 2008, 238-265
- Tirole, J. (1986), "Hierarchies and Bureaucracies: on the Role of Collusion in Organizations," *Journal of Law, Economics and Organization*, 2(2):181-214.
- (1988), *The Theory of Industrial Organization*, Cambridge, MA: MIT Press.
- Tsumagari M. (1999), "Supervision and Firm Structure," working paper, Department of Economics, Keio University.
- van Zandt, T. (1996) "Decentralized Information Processing in the Theory of Organizations," in *Contemporary Economic Development Reviewed, Volume 4: The Enterprise and its Environment*, Murat Sertel (Ed.), London: Macmillan Press.
- (1997), "Real-Time Hierarchical Resource Allocation," *mimeo*, Department of Economics, Princeton University.
- Vaysman I. (1996), "A Model of Cost-Based Transfer Pricing," *Journal of Accounting Studies*, 1.
- (1998), "A Model of Negotiated Transfer Pricing," *Journal of Accounting and Economics*, 25.
- Williamson, Oliver (1967) "Hierarchical Control and Optimal Firm Size," *Journal of Political Economy* 123-138.
- (1985) *The Economic Institutions of Capitalism* New York, Free Press.